TRABAJO FIN DE GRADO

CURSO ACADÉMICO 2023/2024

MANUAL DE MÍNERIA DE DATOS CON EL PAQUETE ESTADÍSTICO R

UNIVERSIDAD MIGUEL HERNÁNDEZ

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN ESTADÍSTICA EMPRESARIAL

ALUMNA: **MARINA RUIZ MEDINA**

TUTORA: **MERCEDES LANDETE RUIZ**

INDICE

1. INTRODUCCIÓN.

La minería de datos puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, es decir, es un proceso de exploración y análisis de grandes conjuntos de datos para descubrir patrones, tendencias y relaciones significativas que puedan utilizarse para tomar decisiones informadas y predecir resultados futuros. Es una disciplina que combina técnicas y herramientas de la estadística, la inteligencia artificial, la informática y las matemáticas.

En términos más simples, la minería de datos implica buscar información valiosa y comprensible dentro de grandes cantidades de datos. Este proceso se lleva a cabo a través de varias etapas, que incluyen la selección y preparación de los datos, la exploración y visualización de los mismos, la construcción de modelos predictivos o descriptivos, y la interpretación y evaluación de los resultados obtenidos.

El análisis multivariante de datos es una técnica dentro del campo de la minería de datos que se utiliza para explorar y analizar conjuntos de datos que contienen múltiples variables interrelacionadas. Este análisis considera simultáneamente varias variables para comprender mejor las relaciones complejas entre ellas.

Esta técnica es especialmente útil cuando se trabaja con conjuntos de datos complejos y ricos en información, donde las variables pueden estar correlacionadas entre sí y pueden influirse mutuamente. Las técnicas estadísticas multivariantes más comunes se pueden clasificar según varios criterios.

-Clasificación por objetivos

a) Reducción de la dimensión:

1. Componentes Principales

2. Análisis Factorial

3. Análisis de Correspondencias

4. Escalamiento Multidimensional

b) Clasificación en grupos:

5. Análisis Clúster

-Clasificación por tipo de variable

a) Variables cuantitativas: 1. Componentes Principales y 2. Análisis Factorial

b) Variables cualitativas: 3. Análisis de Correspondencias

c) Variables cualitativas ordinales: 4. Escalamiento Multidimensional

d) Variables de cualquier tipo: 5. Análisis Clúster

En este trabajo vamos a resolver problemas de las diferentes técnicas de análisis multivariante, estás técnicas incluyen:

- Análisis de componentes principales (ACP): Una técnica utilizada para reducir la dimensionalidad de un conjunto de datos, identificando las direcciones en las que los datos varían más y proyectándolos en un nuevo espacio de menor dimensión.

- Análisis factorial: Similar al ACP, el análisis factorial busca identificar las variables latentes o subyacentes que explican la estructura de correlación observada entre las variables observadas.

- Análisis de correspondencias: Es una técnica descriptiva o exploratoria cuyo objetivo es resumir una gran cantidad de datos en un número reducido de dimensiones, con la menor pérdida de información posible. Su objetivo es similar al de los métodos anteriores, salvo que en el caso del análisis de correspondencias el método se aplica sobre variables categóricas u ordinales. El análisis de correspondencias simples se utiliza a menudo en la representación de datos que se pueden presentar en forma de tablas de contingencia de dos variables.

- Escalamiento: este capítulo no se dedica a una sola técnica sino a un conjunto de ellas. Se llama 'Escalamiento' a cualquier técnica que construye una escala de medida. Se dice que el 'Escalamiento' sirve para ilustrar opiniones.

- Análisis de conglomerados (clustering): Agrupa los datos en grupos o "clústeres" basados en la similitud entre las observaciones, lo que permite identificar patrones o segmentos dentro de los datos.

Estas son solo algunas de las técnicas de análisis multivariante que vamos a ver. En general, el análisis multivariante es una herramienta poderosa para explorar y comprender la complejidad de los datos al considerar las relaciones entre múltiples variables simultáneamente.

Estas técnicas se van a desarrollar con el paquete estadístico R y se van a comparar los resultados con SPSS. R y SPSS son dos herramientas ampliamente utilizadas en el análisis de datos y la estadística, cada una con sus propias características y ventajas:

R: es un lenguaje de programación y un entorno de software estadístico de código abierto. Es altamente flexible y personalizable, lo que permite a los usuarios escribir y ejecutar sus propios scripts para análisis de datos. Cuenta con una amplia variedad de paquetes y extensiones desarrollados por la comunidad de usuarios, que cubren prácticamente todas las técnicas estadísticas imaginables. Es gratuito y de código abierto, lo que significa que cualquiera puede descargarlo, usarlo y contribuir a su desarrollo.

SPSS (Statistical Package for the Social Sciences): es un software comercial desarrollado por IBM que se utiliza principalmente para el análisis de datos en ciencias sociales y empresariales. Tiene una interfaz gráfica de usuario intuitiva que facilita la realización de análisis estadísticos sin necesidad de programación. Ofrece una amplia gama de funciones estadísticas y herramientas de visualización de datos. Es ampliamente utilizado en entornos académicos, de investigación y empresariales debido a su facilidad de uso y a la disponibilidad de soporte técnico.

En resumen, R es preferido por su flexibilidad y su capacidad para realizar análisis complejos y personalizados, mientras que SPSS es más popular entre aquellos que prefieren una interfaz gráfica de usuario y no tienen experiencia en programación. La elección entre R y SPSS a menudo depende de las necesidades específicas del usuario, el nivel de experiencia en programación y el presupuesto disponible.

2. COMPONENTES PRINCIPALES.

Componentes principales es una técnica estadística de reducción de las dimensiones. Es decir, para un conjunto de datos con multitud de variables, su objetivo es el de reducir a un menor número de componentes perdiendo la menor cantidad de información posible. Por lo tanto, el objetivo principal de este análisis en Componentes Principales (CP) es encontrar un conjunto de nuevas variables que sean combinaciones lineales de las variables originales y que capturen la mayor cantidad posible de la variabilidad presente en los datos. Además, a las nuevas variables se les llama componentes principales. Existen dos requisitos, que las variables sean cuantitativas y que exista correlación significativa entre las variables, no es necesaria la normalidad de los datos.

Características:

1. Las CP han de explicar un porcentaje razonable de la varianza global.

2. La extracción de CP se suele realizar sobre variables tipificadas.

3. La representación gráfica de los individuos sobre las CP indica similitud o diferencia de los mismos. Son más similares los individuos más cercanos.

4. Las CP son cuantitativas.

5. El valor de la prueba de Esfericidad de Barlett debe tener una significación inferior a 0.1

6. El análisis es mejor cuanto mayor es el valor KMO.

7. Las comunalidades tras la extracción deben ser superiores a 0.7

En resumen, CP es una herramienta poderosa para reducir la dimensionalidad de los datos mientras se conserva la mayor cantidad posible de información importante, lo que facilita el análisis y la interpretación del conjuntos de datos. Existen varios modos de decidir el número de CP pero nosotros usaremos todas las CP con valor propio mayor que la unidad y complementaremos con las necesarias hasta lograr un 85% de varianza explicada.

2.1. Ejercicio 1. Consideramos el fichero EMPRESAS que contiene información sobre empresas por países y sectores de actividad. Realiza un análisis en Componentes Principales de todas las variables del fichero con la finalidad de reducirlas a un conjunto menor de variables con la menor pérdida de información posible.

Para poder importar los datos debemos instalar el paquete "haven" en R que se utiliza para leer y escribir datos en formatos de software estadísticos como SAS, SPSS y Stata. Esto es especialmente útil cuando se trabaja en un entorno donde se necesita intercambiar datos entre diferentes programas de análisis estadístico. Leemos los datos:

```
#install.packages('haven')
library('haven')
```

```
## Warning: package 'haven' was built under R version 4.2.3
```

```
empresas<-as.data.frame(read_sav("Empresas.sav"))
```

Análisis descriptivo básico

```
summary(empresas)
```

```
##       ID                  AGR              MIN              MAN
##  Length:26          Min.   : 2.70   Min.   :0.100   Min.   : 7.90
##  Class :character   1st Qu.: 7.70   1st Qu.:0.525   1st Qu.:23.00
##  Mode  :character   Median :14.45   Median :0.950   Median :27.55
##                     Mean   :19.13   Mean   :1.254   Mean   :27.01
##                     3rd Qu.:23.68   3rd Qu.:1.800   3rd Qu.:30.20
##                     Max.   :66.80   Max.   :3.100   Max.   :41.20
##       CEN              CON              SER              BAN
##  Min.   :0.1000   Min.   : 2.800   Min.   : 5.20   Min.   : 0.500
##  1st Qu.:0.6000   1st Qu.: 7.525   1st Qu.: 9.25   1st Qu.: 1.225
##  Median :0.8500   Median : 8.350   Median :14.40   Median : 4.650
##  Mean   :0.9077   Mean   : 8.165   Mean   :12.96   Mean   : 4.000
##  3rd Qu.:1.1750   3rd Qu.: 8.975   3rd Qu.:16.88   3rd Qu.: 5.925
##  Max.   :1.9000   Max.   :11.500   Max.   :19.10   Max.   :11.300
##      SECSER            TC
##  Min.   : 5.30   Min.   :3.200
##  1st Qu.:16.25   1st Qu.:5.700
##  Median :19.65   Median :6.700
##  Mean   :20.02   Mean   :6.546
##  3rd Qu.:24.12   3rd Qu.:7.075
##  Max.   :32.40   Max.   :9.400
```

En el análisis descriptivo básico observamos diez variables: ID (País), AGR (Agricultura), MIN (Minería), MAN (Manufactura), CEN (Centrales de energía), CON (Construcción), SER (Servicios a empresas), BAN (Bancos), SECSER (Sector Servicios) y TC (Transporte y comunicaciones). La primera respectivamente se trata de una variable cualitativa que cuenta con la cantidad de 26 paises mientras que las demás son cuantitativas y todas se encuentran en el mismo orden de magnitud.

Matriz de correlaciones

Supongamos que tienes n variables $X_1, X_2, \ldots, X_n$. La matriz de correlaciones se vería así:

$$\begin{bmatrix} 1 & \text{Corr}(X_1, X_2) & \ldots & \text{Corr}(X_1, X_n) \\ \text{Corr}(X_2, X_1) & 1 & \ldots & \text{Corr}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Corr}(X_n, X_1) & \text{Corr}(X_n, X_2) & \ldots & 1 \end{bmatrix}$$

Donde:

- $\text{Corr}(X_i, X_j)$ es el coeficiente de correlación entre las variables X_i y X_j.
- La diagonal principal contiene unos (ya que la correlación de una variable consigo misma es siempre 1).

Una matriz de correlaciones es una matriz cuadrada que contiene los coeficientes de correlación entre todas las posibles combinaciones de pares de variables en un conjunto de datos.

Si utilizamos la función "*cor(empresas)*" nos va a dar error porque hay una columna con valores no numéricos, por lo tanto, primero debemos asignar a cada fila el nombre id y luego eliminaremos la columna que contiene los nombres:

Para ellos necesitaremos instalar el paquete "corrplot". El paquete corrplot es un paquete muy flexible, que permite crear una amplia variedad de correlogramas con una sola función. Se recomienda ejecutar "?corrplot" para obtener detalles adicionales. Ten en cuenta que en esta función se debe pasar la matriz de correlación, en lugar de las variables.

```
rownames(empresas)<-empresas$ID
empresas <- empresas[ ,!colnames(empresas)=="ID"]
cor(empresas) #matriz de correlaciones
```

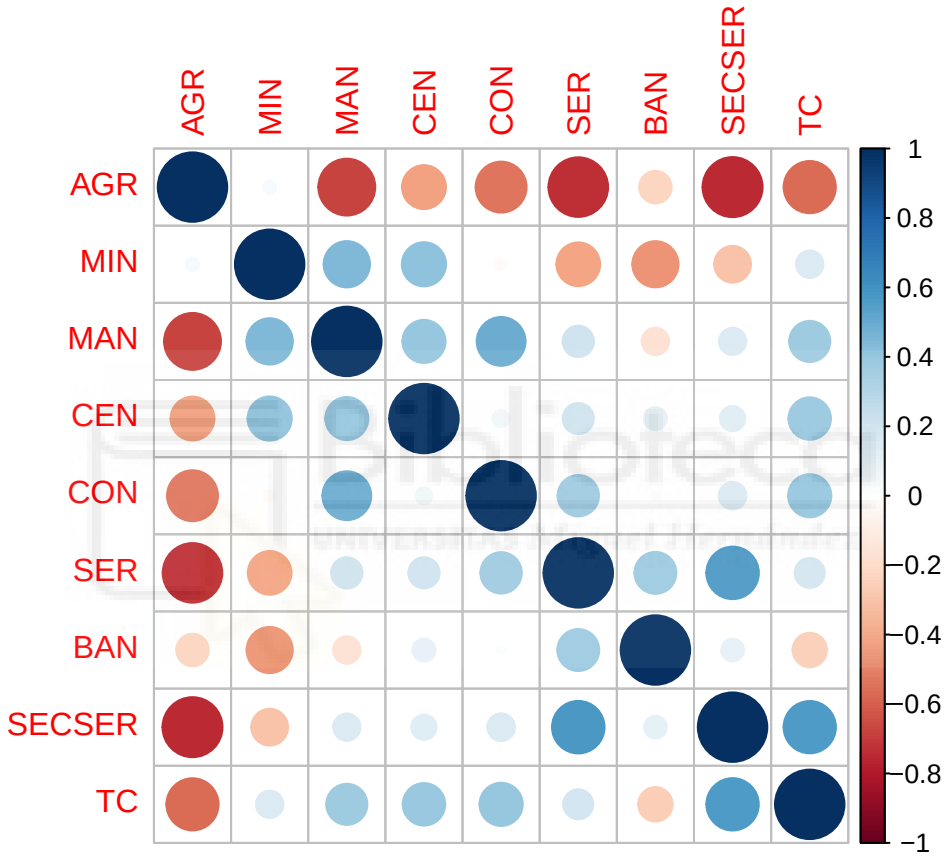
```
##                AGR         MIN        MAN         CEN         CON         SER
## AGR     1.00000000  0.03579884 -0.6710976 -0.40005113 -0.53832522 -0.7369805
## MIN     0.03579884  1.00000000  0.4451960  0.40545524 -0.02559781 -0.3965646
## MAN    -0.67109759  0.44519601  1.0000000  0.38534593  0.49447949  0.2038263
## CEN    -0.40005113  0.40545524  0.3853459  1.00000000  0.05988883  0.2019066
## CON    -0.53832522 -0.02559781  0.4944795  0.05988883  1.00000000  0.3560216
## SER    -0.73698054 -0.39656456  0.2038263  0.20190661  0.35602160  1.0000000
## BAN    -0.21983645 -0.44268311 -0.1558288  0.10986158  0.01628255  0.3655553
## SECSER -0.74679001 -0.28101212  0.1541714  0.13241132  0.15824309  0.5721728
## TC     -0.56492047  0.15662892  0.3506925  0.37523116  0.38766214  0.1875543
##                BAN     SECSER         TC
## AGR    -0.21983645 -0.7467900 -0.5649205
## MIN    -0.44268311 -0.2810121  0.1566289
## MAN    -0.15582884  0.1541714  0.3506925
## CEN     0.10986158  0.1324113  0.3752312
## CON     0.01628255  0.1582431  0.3876621
## SER     0.36555529  0.5721728  0.1875543
## BAN     1.00000000  0.1076403 -0.2459257
## SECSER  0.10764028  1.0000000  0.5678669
## TC     -0.24592567  0.5678669  1.0000000
```

```
#install.packages("corrplot")
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(empresas))
```



Ahora ha funcionado bien porque hemos eliminado la columna que contiene los nombres de los paises (variable cualitativa). En un Análisis en Componentes Principales sólo se pueden utilizar variables cuantitativas. Si tuvieramos un ejercicio con una base de datos en la que hubieran varias variables cualitativas debemos eliminarlas o seleccionar solamente las cuantitativas.

Fijándonos en la matriz de correlaciones podemos obtener las siguientes conclusiones: se observa que existen puntos muy grandes y oscuros, lo que significa que hay variables muy correlacionadas. Se asemeja a la matriz identidad, ya que fuera de la diagonal no hay puntos muy grandes. Además, la correlación entre (sector servicios y centrales de energía) es de 0,132 mientras que la correlación entre (transportes y comunicaciones y manufactura) es de 0,351.

Tipificamos los datos

```
G<-cov.wt(empresas, method='ML')$center
V<-cov.wt(empresas, method='ML')$cov
escala<-diag(V)
Z<-scale(empresas, center=G, scale=sqrt(escala))
head(Z)
```

```
##                    AGR        MIN         MAN         CEN         CON       SER
## Bélgica     -1.0384465 -0.3719974  0.08619551 -0.02085144  0.02145187 1.3690936
## Dinamarca   -0.6514259 -1.2130350 -0.75784885 -0.83405766  0.08342395 0.3660632
## Francia     -0.5464711 -0.4771271  0.07164302 -0.02085144  0.45525640 0.8564336
## Alemania O -0.8154177  0.0485214  1.27949961 -0.02085144 -0.53629681 0.3214841
## Irlanda      0.2669282 -0.2668677 -0.91792623  1.06342351 -0.41235266 0.8564336
## Italia      -0.2119279 -0.6873865  0.08619551 -1.10512639  1.13694923 1.1461980
##                    BAN     SECSER         TC
## Bélgica      0.7994005 0.98208220  0.4792023
## Dinamarca    0.9084097 1.81828787  0.4059125
## Francia      0.7267278 0.38479244 -0.6201441
## Alemania O   0.3633639 0.33999571 -0.3269851
## Irlanda     -0.4360367 0.11601205 -0.3269851
## Italia      -0.8720733 0.01148634 -0.6201441
```

Obtención de las Componentes Principales

```
CP<-prcomp(Z)

summary(CP)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5    PC6     PC7
## Standard deviation     1.9044 1.4884 1.0691 1.0170 0.75163 0.6315 0.48455
## Proportion of Variance 0.3875 0.2367 0.1221 0.1105 0.06036 0.0426 0.02508
## Cumulative Proportion  0.3875 0.6241 0.7462 0.8568 0.91711 0.9597 0.98480
##                           PC8      PC9
## Standard deviation     0.3772 0.006888
## Proportion of Variance 0.0152 0.000010
## Cumulative Proportion  1.0000 1.000000
```

```
summary(CP)$importance[2,] #Proporción de varianza explicada.
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## 0.38746 0.23669 0.12211 0.11050 0.06036 0.04260 0.02508 0.01520 0.00001
```

```
summary(CP)$importance[3,]*100 #Porcentaje de varianza acumulada.
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
##  38.746  62.415  74.625  85.675  91.711  95.971  98.480  99.999 100.000
```

El porcentaje de varianza explicado ha de ser superior al 80%, por lo tanto, nos quedamos con cuatro factores que explican el 85.675% de la varianza, para obtener este resultado nos fijamos en 'Cumulative Proportion' en R, mientras que en SPSS nos fijamos en 'Autovalores iniciales % acumulado'. La primera componente acumula el 38.746%, la segunda el 62.415% y la tercera el 74.625%.

La tercera componente captura 'casi toda' la variabilidad de los datos. Esto quiere decir que podríamos reducir las 9 variables originales a tres variables (componentes principales) manteniendo (prácticamente) constante la cantidad de información disponible con respecto al conjunto de datos originales.

Para hacer una representación gráfica de los individuos sobre las Componentes Principales (CP), utilizariamos el siguiente código:

```
# biplot(CP, scale=0, col=c("red","blue"))
```

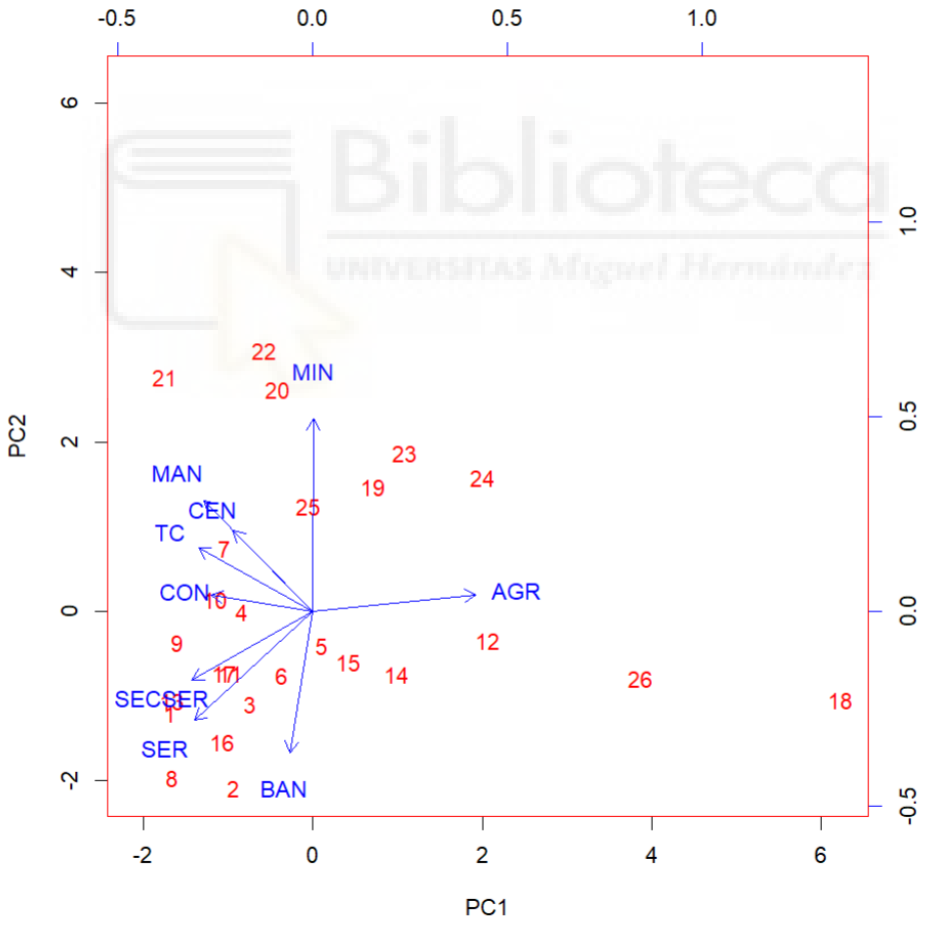


Figure 1: "Representación gráfica de los individuos sobre las CP"

El país que invierte en Mineria (MIN), como, por ejemplo, Hungría, no invierte en Bancos (BAN), como, por ejemplo, Dinamarca. Los sectores que más se diferencian son AGR (Agricultura) de TC (Transporte y comunicaciones) y CON (Construcción), van en sentido contrario. El país que invierte en Manufactura también invierte en Centrales de energía, como, por ejemplo, Luxemburgo, sin embargo, este no invierte en agricultura, como, por ejemplo, Grecia.

Los sectores de actividad (más relacionadas, man (Manufactura) y cen (Centrales de energía), misma dirección), también serian actividades relacionadas SER (Servicios a empresas) y SECSER (Sector servicios).

A continuación se calcula la matriz de varianzas y covarianzas entre cada par de variables. Como en este ejemplo hay NUEVE variables, el resultado es una matriz simétrica 9x9.

```
matriz_cov <- cov(empresas);matriz_cov
```

```
##                AGR         MIN        MAN         CEN          CON         SER
## AGR    241.6958154  0.53987692 -73.113846 -2.33984615 -13.77209231 -52.4210462
## MIN      0.5398769  0.94098462   3.026369  0.14796923  -0.04086154  -1.7600308
## MAN    -73.1138462  3.02636923  49.108738  1.01593846   5.70227692   6.5351385
## CEN     -2.3398462  0.14796923   1.015938  0.14153846   0.03707692   0.3475385
## CON    -13.7720923 -0.04086154   5.702277  0.03707692   2.70795385   2.6804769
## SER    -52.4210462 -1.76003077   6.535138  0.34753846   2.68047692  20.9329385
## BAN     -9.5920000 -1.20520000  -3.064800  0.11600000   0.07520000   4.6940000
## SECSER -79.2911385 -1.86169231   7.378615  0.34021538   1.77843077  17.8786154
## TC     -12.2206769  0.21141538   3.419631  0.19643077   0.88766154   1.1940308
##            BAN      SECSER         TC
## AGR    -9.5920 -79.2911385 -12.2206769
## MIN    -1.2052  -1.8616923   0.2114154
## MAN    -3.0648   7.3786154   3.4196308
## CEN     0.1160   0.3402154   0.1964308
## CON     0.0752   1.7784308   0.8876615
## SER     4.6940  17.8786154   1.1940308
## BAN     7.8768   2.0632000  -0.9604000
## SECSER  2.0632  46.6426462   5.3964923
## TC     -0.9604   5.3964923   1.9361846
```

Una matriz de varianzas y covarianzas es una matriz cuadrada que contiene las varianzas de las variables en la diagonal principal y las covarianzas entre pares de variables en las posiciones fuera de la diagonal principal. Esta matriz proporciona información sobre la dispersión de cada variable individualmente y cómo están relacionadas entre sí en términos de covarianza. La variable agricultura tiene una varianza de 241,696. Además, tiene una covarianza con la variable construcción de -13,772. La variable bancos tiene una varianza de 7,877 y una covarianza con la variable minería de -1,205.

Dado que la matriz de covarianzas es cuadrada, se pueden obtener sus correspondientes eigenvectors y eigenvalues.

```
eigen <- eigen(matriz_cov);eigen
```

```
## eigen() decomposition
## $values
## [1] 3.034581e+02 4.370166e+01 1.520735e+01 5.639360e+00 2.443399e+00
## [6] 1.046033e+00 4.208472e-01 6.492666e-02 1.911981e-03
##
## $vectors
##               [,1]         [,2]         [,3]        [,4]         [,5]
##  [1,]  0.891758406  0.006826746  0.118466699 -0.09676712  0.180043781
##  [2,]  0.001922618 -0.092347069  0.079379068 -0.01015633 -0.001121643
##  [3,] -0.271271411 -0.770269221  0.184678991 -0.01040077  0.335999746
##  [4,] -0.008388285 -0.012015922 -0.006768322  0.01814178 -0.002459689
##  [5,] -0.049594016 -0.068988571 -0.077312766 -0.08292614 -0.724262390
##  [6,] -0.191798409  0.234416513 -0.579612752 -0.60760858  0.265863007
##  [7,] -0.031128614  0.130082403 -0.469969939  0.78119316  0.121062046
##  [8,] -0.298046310  0.566777401  0.597745181  0.04833726  0.235915950
##  [9,] -0.045364280  0.009888386  0.159415225 -0.03783527 -0.434890328
##              [,6]         [,7]         [,8]      [,9]
##  [1,] -0.15262561  0.091621401  0.068678066 0.3354111
##  [2,]  0.45636121 -0.766470364  0.290464275 0.3239614
##  [3,] -0.20093094  0.161983468  0.074117735 0.3374633
##  [4,]  0.23086414 -0.062936752 -0.909183254 0.3398982
##  [5,] -0.55835746 -0.194294560 -0.004457936 0.3253270
##  [6,] -0.02157242  0.087935421  0.104435658 0.3366529
##  [7,] -0.05528170  0.079976564  0.122754675 0.3343621
##  [8,] -0.24786088  0.004543731  0.052137300 0.3323638
##  [9,]  0.54593853  0.567476088  0.223813567 0.3342147
```

Los valores propios (también conocidos como autovalores) y los vectores propios (también conocidos como autovectores) son herramientas importantes para comprender cómo las transformaciones lineales afectan a los vectores en diferentes direcciones. En el análisis de datos, los valores y vectores propios son utilizados en técnicas como el Análisis de Componentes Principales (CP).

Test de Barlett. Test de esfericidad de Bartlett: este test prueba la hipótesis nula de que las variables están incorrelacionadas, es decir, evalúa si la matriz de correlaciones no es una matriz de identidad, aquella en la que no existe relación entre las variables. Se rechaza cuando el nivel de significación es menor al 5%.

Ho: la matriz de correlaciones no es la matriz identidad (no existe relación entre las variables)

H1: en caso contrario

El paquete "psych" en R es una herramienta diseñada para análisis estadísticos. Ofrece una amplia gama de funciones para realizar análisis de datos.

```
#install.packages('psych')
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.2.3
```

```
KMO(cor(empresas))
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(empresas))
## Overall MSA =  0.13
## MSA for each item =
##    AGR    MIN    MAN    CEN    CON    SER    BAN SECSER     TC
##   0.24   0.10   0.14   0.10   0.10   0.15   0.06   0.15   0.14
```

```
bartlett.test(empresas)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  empresas
## Bartlett's K-squared = 382.44, df = 8, p-value < 2.2e-16
```

Por lo tanto, el análisis en Componentes Principales (CP) es adecuado porque la prueba de esfericidad de Bartlett nos da un p-value $<$ 2.2e-16 $<$ (0.05), rechazo H0, si existe relación entre las variables, lo cual nos indica que la matriz de datos SI es válida para continuar con el proceso. El valor KMO es 0.13 (el análisis es mejor cuanto mayor es su valor).

Comunalidades.

En el contexto del Análisis de Componentes Principales (CP), las comunalidades representan la proporción de la varianza de cada variable que es explicada por los componentes principales. En otras palabras, las comunalidades indican qué tan bien los componentes principales capturan la información contenida en las variables originales. Interpretación: Una comunalidad alta (cercana a 1) indica que una gran parte de la varianza de la variable se explica por los componentes principales, mientras que una comunalidad baja sugiere que la variable no está bien representada por los componentes seleccionados.

```
factores<-fa(cor(Z), nfactors=4, fm='minres', rotate='none')
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
## ultra-Heywood case was detected.  Examine the results carefully
```

```
factores$communalities
```

```
##       AGR       MIN       MAN       CEN       CON       SER       BAN    SECSER
## 0.9950000 0.7104734 0.8606482 0.9950000 0.4076277 0.6880879 0.5385504 0.8417338
##        TC
## 0.6670115
```

Interpretación de los resultados

Después de calcular las comunalidades, cada valor nos indica la proporción de la varianza de cada variable original que es explicada por los componentes principales seleccionados. Por ejemplo, la variable agricultura tiene una comunalidad de 0.995, esto significa que el 99.5% de la variabilidad de la variable se explica por las componentes principales, mientras que la variable servicios tiene una comunalidad de 0.688, lo que significa que el 68.8% de la variabilidad de esta variable se explica por las componentes principales.

En resumen, las comunalidades son la información que retienen las variables y deben ser superiores a 0.7, por lo que deberíamos eliminar la variable 'CON' con una comunalidad de 0.407 y 'BAN' con una comunalidad 0.538. Las comunalidades son una herramienta esencial para evaluar la calidad de la representación de las variables originales en el espacio reducido de los componentes principales.

2.2. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Analizar - Reducción de dimensiones - Análisis Factorial - Pasamos todas las variables cuantitativas, en este caso todas menos país - En descriptivos seleccionamos KMO y prueba de esfericidad de Barlett, Matriz de covarianzas - En extracción seleccionamos el númer de factores que queremos extraer, en nuestro caso seleccionamos 4 factores - Aceptar

Obtenemos los siguientes resultados:

En la primera tabla que nos aparece nos fijamos en autovalores iniciales en el % acumulado, como bien hemos dicho anteriormente el porcentaje de varianza explicado ha de ser superior al 80%, por lo tanto, nos quedamos con cuatro factores que explican el 85.675% de la varianza.

Varianza total explicada

	Autovalores iniciales			Sumas de cargas al cuadrado de la extracción		
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	3,487	38,746	38,746	3,487	38,746	38,746
2	2,130	23,669	62,415	2,130	23,669	62,415
3	1,099	12,211	74,625	1,099	12,211	74,625
4	,994	11,050	85,675	,994	11,050	85,675
5	,543	6,036	91,711			
6	,383	4,260	95,971			
7	,226	2,508	98,480			
8	,137	1,520	99,999			
9	4,563E-5	,001	100,000			

Método de extracción: análisis de componentes principales.

Figure 2: "COMPONENTES PRINCIPALES"

Matriz de correlaciones

		agricultura	minería	manufactura	centrales de energía	construcción	servicios a empresas	bancos	sector servicios	transporte y comunicaciones
Correlación	agricultura	1,000	,036	-,671	-,400	-,538	-,737	-,220	-,747	-,565
	minería	,036	1,000	,445	,405	-,026	-,397	-,443	-,281	,157
	manufactura	-,671	,445	1,000	,385	,494	,204	-,156	,154	,351
	centrales de energía	-,400	,405	,385	1,000	,060	,202	,110	,132	,375
	construcción	-,538	-,026	,494	,060	1,000	,356	,016	,158	,388
	servicios a empresas	-,737	-,397	,204	,202	,356	1,000	,366	,572	,188
	bancos	-,220	-,443	-,156	,110	,016	,366	1,000	,108	-,246
	sector servicios	-,747	-,281	,154	,132	,158	,572	,108	1,000	,568
	transporte y comunicaciones	-,565	,157	,351	,375	,388	,188	-,246	,568	1,000

Figure 3: "MATRIZ DE CORRELACIONES"

Matriz de covarianzas

	agricultura	minería	manufactura	centrales de energía	construcción	servicios a empresas	bancos	sector servicios	transporte y comunicaciones
agricultura	241,696	,540	-73,114	-2,340	-13,772	-52,421	-9,592	-79,291	-12,221
minería	,540	,941	3,026	,148	-,041	-1,760	-1,205	-1,862	,211
manufactura	-73,114	3,026	49,109	1,016	5,702	6,535	-3,065	7,379	3,420
centrales de energía	-2,340	,148	1,016	,142	,037	,348	,116	,340	,196
construcción	-13,772	-,041	5,702	,037	2,708	2,680	,075	1,778	,888
servicios a empresas	-52,421	-1,760	6,535	,348	2,680	20,933	4,694	17,879	1,194
bancos	-9,592	-1,205	-3,065	,116	,075	4,694	7,877	2,063	-,960
sector servicios	-79,291	-1,862	7,379	,340	1,778	17,879	2,063	46,643	5,396
transporte y comunicaciones	-12,221	,211	3,420	,196	,888	1,194	-,960	5,396	1,936

Figure 4: "MATRIZ DE VARIANZAS Y COVARIANZAS"

Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,134
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	274,053
	gl	36
	Sig.	<,001

Figure 5: "TEST DE BARLETT Y KMO"

Comunalidades

	Inicial	Extracción
agricultura	1,000	,966
minería	1,000	,862
manufactura	1,000	,834
centrales de energía	1,000	,873
construcción	1,000	,844
servicios a empresas	1,000	,779
bancos	1,000	,840
sector servicios	1,000	,904
transporte y comunicaciones	1,000	,810

Método de extracción: análisis de componentes principales.

Figure 6: "COMUNALIDADES"

Si nos fijamos en ambas salidas, tanto en la que hemos visto antes de R como en la de SPSS, observamos que las comunalidades son parecidas, algunas coinciden, en cambio otras son diferentes. Por ejemplo, la variable 'CON' (Construcción), tiene una comunalidad en R de 0.407, mientras que en SPSS tiene 0.844, en R llegamos a la conclusión de que deberíamos eliminarla, mientras que en SPSS no eliminariamos ninguna variable.

3. ANÁLISIS FACTORIAL.

El análisis factorial es una técnica estadística utilizada para explorar las relaciones entre un conjunto de variables observadas y reducir la dimensionalidad de los datos, el objetivo es similar al de CP porque también busca un número de nuevas variables, menor al número de variables originales que representen del mejor modo posible a las originales. En este caso a las nuevas variables se les llama Factores.

El análisis factorial y el análisis de componentes principales (CP) son dos técnicas de reducción de dimensionalidad ampliamente utilizadas en estadística multivariada, pero difieren en su enfoque y objetivos. La principal diferencia entre ambos es:

El análisis factorial busca identificar factores que expliquen la covariación entre un conjunto de variables observadas, mientras que el análisis de componentes principales, en cambio, busca describir la estructura de covarianza (o correlación) entre las variables observadas utilizando un conjunto reducido de componentes que explican la mayor parte de la variabilidad en los datos.

En resumen, aunque el análisis factorial y el análisis de componentes principales comparten similitudes en su aplicación de reducción de dimensionalidad, difieren en sus objetivos e interpretación. La elección entre ambos métodos depende del contexto específico de la investigación y de los objetivos del análisis.

Características:

1. **Las variables son cuantitativas.**

2. **El análisis es mejor cuanto mayor es el valor KMO.**

3. **Las comunalidades tras la extracción deben ser superiores a 0.6**

4. **El porcentaje de varianza explicada ha de ser superior al 70%.**

5. **La extracción se suele realizar sobre variables tipificadas.**

6. **El valor de la prueba de Esfericidad de Barlett debe tener una significación inferior a 0.1**

3.1. Ejercicio 2. Una empresa especializada en el diseño de automóviles de turismo desea estudiar cuales son los deseos del público que compra automóviles. Para ello diseña una encuesta con 10 preguntas donde se le pide a cada uno de los 20 encuestados que valore de 1 a 5 si una característica es o no muy importante. Los encuestados deben contestar con un 5 si la característica es muy importante, un 4 si es importante, un 3 si tiene regular importancia, un 2 si es poco importante y un 1 si no es nada importante. Las 10 características (V1 a V10) a valorar son: precio, financiación, consumo, combustible, seguridad, confort, capacidad, prestaciones, modernidad y aerodinámica. El fichero 6-2.sav recoge los datos. Realiza un análisis factorial que resuma correctamente la información.

Cargamos las librerias:

El paquete "nFactors" en R se utiliza para determinar el número óptimo de factores a extraer en análisis factorial (AF).

```
#install.packages('nFactors')
library('nFactors')
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'nFactors'
```

```
## The following object is masked from 'package:lattice':
##
##     parallel
```

```
#install.packages('psych')
library('psych')
#install.packages('corrplot')
library('corrplot')
```

Leemos los datos:

```
#install.packages('haven')
library('haven')
estudio <- as.data.frame(read_sav("6-2.sav"))
summary(estudio)
```

```
##        V1              V2              V3              V4              V5
##  Min.   :1.0   Min.   :1.0   Min.   :1.00   Min.   :1.0   Min.   :2.0
##  1st Qu.:3.0   1st Qu.:2.0   1st Qu.:2.75   1st Qu.:1.0   1st Qu.:3.0
##  Median :4.0   Median :4.0   Median :4.00   Median :3.0   Median :4.0
##  Mean   :3.7   Mean   :3.4   Mean   :3.50   Mean   :2.8   Mean   :3.7
##  3rd Qu.:5.0   3rd Qu.:5.0   3rd Qu.:4.00   3rd Qu.:4.0   3rd Qu.:4.0
##  Max.   :5.0   Max.   :5.0   Max.   :5.00   Max.   :5.0   Max.   :5.0
```

```
##        V6              V7              V8              V9              V10
##  Min.   :2.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:2.75   1st Qu.:3.00   1st Qu.:2.00   1st Qu.:1.75   1st Qu.:1.75
##  Median :4.00   Median :4.00   Median :3.00   Median :2.50   Median :2.50
##  Mean   :3.70   Mean   :3.65   Mean   :2.85   Mean   :2.80   Mean   :2.65
##  3rd Qu.:5.00   3rd Qu.:4.25   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :5.00   Max.   :5.00   Max.   :5.00   Max.   :5.00   Max.   :5.00
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	4	1	4	3	3	2	4	4	4	4
2	5	5	4	4	3	3	4	1	1	3
3	2	1	3	1	4	2	1	5	4	5
4	1	1	1	1	4	4	2	5	5	4
5	1	1	2	1	5	5	4	3	3	2
6	5	5	5	5	3	3	4	2	2	1
7	4	5	4	4	2	2	5	1	1	1
8	3	2	3	1	4	4	2	5	5	5
9	4	4	4	3	4	4	3	1	1	1
10	5	5	5	5	2	2	3	2	2	2
11	2	2	2	1	5	4	4	3	4	3
12	4	4	5	5	4	5	5	2	1	2
13	3	2	2	1	4	5	4	4	3	3
14	5	5	4	4	5	4	4	1	2	2
15	4	3	3	1	4	4	5	3	4	4
16	5	5	4	4	4	5	4	2	1	1
17	4	4	5	2	4	5	5	4	4	2
18	5	5	4	4	2	2	1	2	2	3
19	3	3	2	2	4	4	5	4	5	4
20	5	5	4	4	4	5	4	3	2	1

Figure 7: “BASE DE DATOS”

Podemos observar que en la base de datos tenemos 10 variables que son: precio (V1),financiación (V2), consumo (V3), combustible (V4), seguridad (V5), confort (V6), capacidad (V7), prestaciones (V8), modernidad (V9) y aerodinámica (V10). Vemos que todas las variables son cuantitativas por lo que no haría falta modificar la base datos, además se encuentran en el mismo orden de magnitud.

Tipificamos los datos

```
G=cov.wt(estudio, method='ML')$center
V=cov.wt(estudio, method='ML')$cov
escala=diag(cov.wt(estudio, method='ML')$cov)
Z=scale(estudio,center=G,scale=sqrt(escala))
Z=as.matrix(Z)
```

Matriz de varianzas y covarianzas

```
matriz_cov<-cov(estudio);matriz_cov
```

```
##             V1         V2         V3         V4         V5          V6
## V1   1.8000000  1.9157895  1.3157895  1.7263158 -0.6210526 -0.30526316
## V2   1.9157895  2.6736842  1.4210526  2.1368421 -0.6631579 -0.13684211
## V3   1.3157895  1.4210526  1.4210526  1.5263158 -0.5263158 -0.31578947
## V4   1.7263158  2.1368421  1.5263158  2.4842105 -0.8000000 -0.48421053
## V5  -0.6210526 -0.6631579 -0.5263158 -0.8000000  0.8526316  0.80000000
## V6  -0.3052632 -0.1368421 -0.3157895 -0.4842105  0.8000000  1.37894737
## V7   0.3631579  0.5157895  0.2894737  0.3473684  0.2052632  0.62631579
## V8  -1.2052632 -1.7789474 -0.9210526 -1.6105263  0.3736842  0.21578947
## V9  -1.2736842 -1.8105263 -1.1052632 -1.8315789  0.4631579  0.09473684
## V10 -0.9000000 -1.5368421 -0.8684211 -1.3894737  0.1526316 -0.37368421
##             V7         V8          V9         V10
## V1   0.3631579 -1.2052632 -1.27368421 -0.9000000
## V2   0.5157895 -1.7789474 -1.81052632 -1.5368421
## V3   0.2894737 -0.9210526 -1.10526316 -0.8684211
## V4   0.3473684 -1.6105263 -1.83157895 -1.3894737
## V5   0.2052632  0.3736842  0.46315789  0.1526316
## V6   0.6263158  0.2157895  0.09473684 -0.3736842
## V7   1.6078947 -0.5289474 -0.33684211 -0.7078947
## V8  -0.5289474  1.9236842  1.81052632  1.3657895
## V9  -0.3368421  1.8105263  2.16842105  1.5578947
## V10 -0.7078947  1.3657895  1.55789474  1.8184211
```

En esta matriz de varianzas y covarianzas podemos observar que la variable (V1) tiene una varianza de 1.80, mientras que la variable (V9) tiene una varianza de 2.16842105, además podemos decir que la variable V2 y V5 tienen una covarianza de -0.6631579, la variable V3 y V7 tienen una covarianza de 0.2894737.

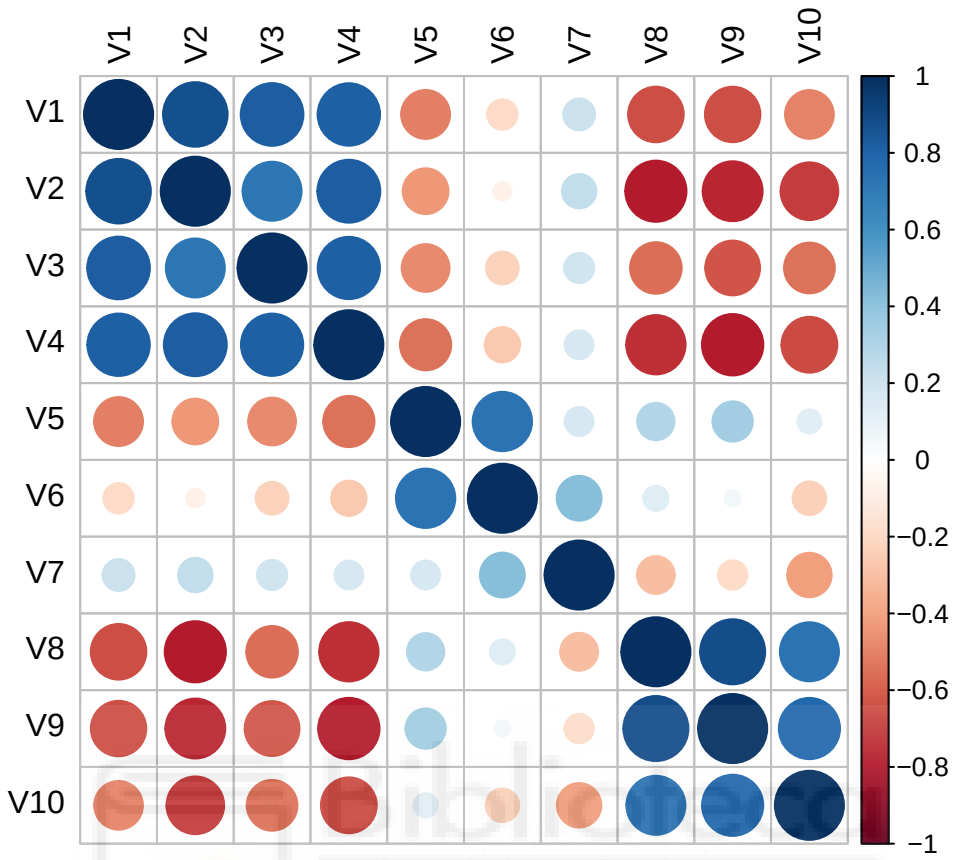
Matriz de correlaciones

```
cor(estudio)
```

```
##             V1          V2         V3         V4         V5          V6
## V1   1.0000000  0.87328595  0.8227068  0.8163752 -0.5013159 -0.19376008
## V2   0.8732860  1.00000000  0.7290378  0.8291310 -0.4392191 -0.07126739
## V3   0.8227068  0.72903777  1.0000000  0.8123536 -0.4781461 -0.22558942
## V4   0.8163752  0.82913105  0.8123536  1.0000000 -0.5496865 -0.26161713
## V5  -0.5013159 -0.43921906 -0.4781461 -0.5496865  1.0000000  0.73779454
## V6  -0.1937601 -0.07126739 -0.2255894 -0.2616171  0.7377945  1.00000000
## V7   0.2134668  0.24876462  0.1915028  0.1738070  0.1753079  0.42062076
## V8  -0.6477072 -0.78440645 -0.5570735 -0.7367273  0.2917811  0.13249196
## V9  -0.6446941 -0.75193098 -0.6296349 -0.7891501  0.3406250  0.05478646
## V10 -0.4974610 -0.69699068 -0.5402292 -0.6537458  0.1225791 -0.23598461
##             V7         V8          V9         V10
## V1   0.2134668 -0.6477072 -0.64469411 -0.4974610
## V2   0.2487646 -0.7844064 -0.75193098 -0.6969907
## V3   0.1915028 -0.5570735 -0.62963492 -0.5402292
## V4   0.1738070 -0.7367273 -0.78915014 -0.6537458
## V5   0.1753079  0.2917811  0.34062503  0.1225791
## V6   0.4206208  0.1324920  0.05478646 -0.2359846
## V7   1.0000000 -0.3007577 -0.18039552 -0.4139927
## V8  -0.3007577  1.0000000  0.88647429  0.7302468
## V9  -0.1803955  0.8864743  1.00000000  0.7845472
## V10 -0.4139927  0.7302468  0.78454720  1.0000000
```

```
corrplot(cor(estudio), tl.col='black',width = 5, height = 5) #Gráfico de las correlaciones
```

```
## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt =
## tl.srt, : "width" is not a graphical parameter

## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt =
## tl.srt, : "height" is not a graphical parameter

## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col =
## tl.col, : "width" is not a graphical parameter

## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col =
## tl.col, : "height" is not a graphical parameter

## Warning in title(title, ...): "width" is not a graphical parameter

## Warning in title(title, ...): "height" is not a graphical parameter
```



Fijándonos en la matriz de correlaciones podemos obtener las siguientes conclusiones: se observa que existen puntos muy grandes y oscuros, lo que significa que hay variables muy correlacionadas. Vemos que no se asemeja a la matriz identidad, ya que fuera de la diagonal existen algunos puntos muy grandes. Además, la correlación entre precio y financiación (V1 y V2) es de 0.87328595 mientras que la correlación entre capacidad y confort (V7 y V6) es de 0.42062076

Test KMO y Test de Barlett.

El test de Barlett nos dice que si el pvalor>0.05, entonces, aceptamos la hipótesis nula.

Ho:las variables no están correlacionadas

H1:las variabels si están correlacionadas

```
KMO(estudio)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = estudio)
## Overall MSA =  0.7
## MSA for each item =
##   V1   V2   V3   V4   V5   V6   V7   V8   V9  V10
## 0.82 0.74 0.84 0.93 0.55 0.32 0.37 0.62 0.68 0.84
```

```
bartlett.test(estudio)
```

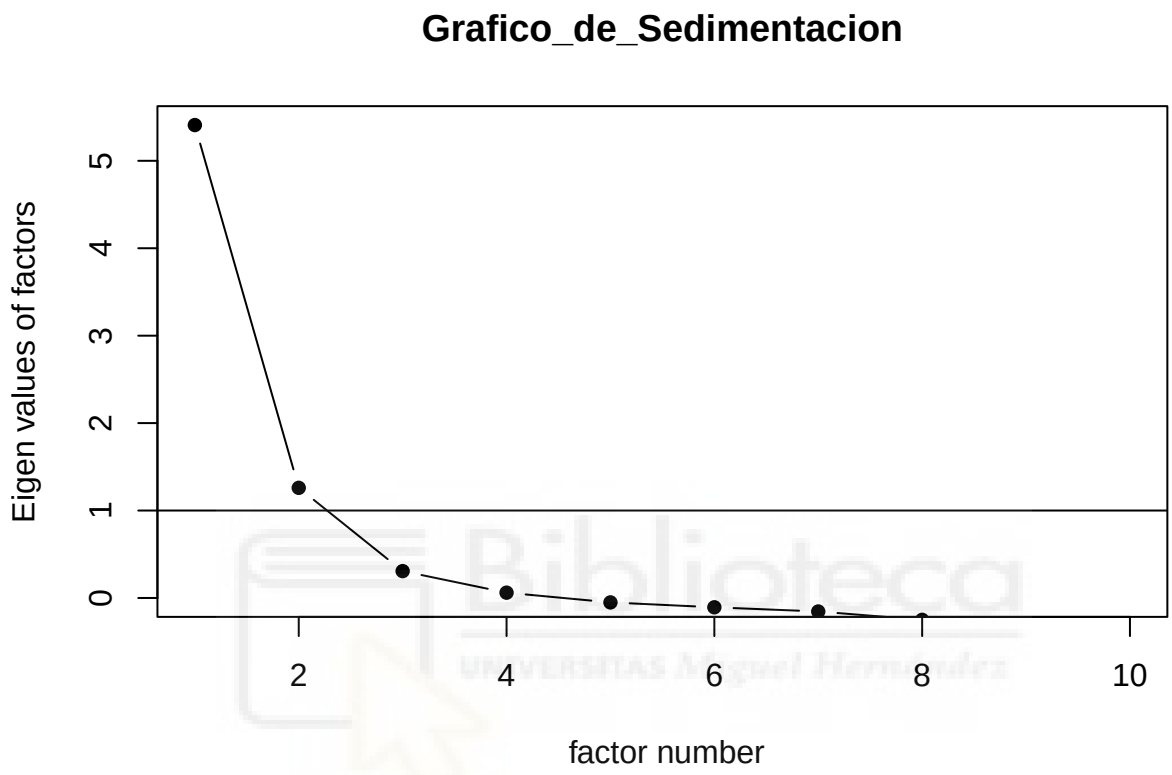
```
##
##  Bartlett test of homogeneity of variances
##
## data:  estudio
## Bartlett's K-squared = 8.4557, df = 9, p-value = 0.489
```

El resultado del KMO es 0.7 lo que nos dice que podemos continuar con el análisis factorial.

Obtenemos un p-value = 0.489>0.05, por lo tanto acepto Ho, lo que quiere decir que las variables no están correlacionadas, lo cual nos indica que la matriz de datos NO es válida para continuar con el proceso, aún así nosotros vamos a continuar.

Determinar el número de factores, para ello utilizamos un Gráfico de sedimentación.

```
scree(estudio, main="Grafico_de_Sedimentacion",pc=FALSE)
```



Observamos dónde corta, siempre nos quedamos con el punto por encima, por lo que el análisis sugiere utilizar 2 factores.

Obtención de los factores. Para ello debemos introducir la matriz con las variables tipificados (cor(Z)), el número de factores, en este caso 2, (nfactors=2) y el método (fm='minres'). Esta función nos da mucha información.

```
factores<-fa(cor(Z), nfactors=2, fm='minres');factores
```

```
## Loading required namespace: GPArotation

## Factor Analysis using method =  minres
## Call: fa(r = cor(Z), nfactors = 2, fm = "minres")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       MR1   MR2   h2   u2 com
## V1   0.82 -0.18 0.74 0.26 1.1
## V2   0.92 -0.02 0.85 0.15 1.0
## V3   0.77 -0.19 0.66 0.34 1.1
```

```
## V4   0.89 -0.20 0.88 0.12 1.1
## V5  -0.33  0.73 0.72 0.28 1.4
## V6   0.04  0.93 0.86 0.14 1.0
## V7   0.37  0.46 0.29 0.71 1.9
## V8  -0.86 -0.06 0.72 0.28 1.0
## V9  -0.88 -0.06 0.76 0.24 1.0
## V10 -0.84 -0.36 0.75 0.25 1.4
##
##                        MR1  MR2
## SS loadings           5.37 1.88
## Proportion Var        0.54 0.19
## Cumulative Var        0.54 0.73
## Proportion Explained  0.74 0.26
## Cumulative Proportion 0.74 1.00
##
##  With factor correlations of
##       MR1   MR2
## MR1  1.00 -0.15
## MR2 -0.15  1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  45  with the objective function =  11.02
## df of  the model are 26  and the objective function was  2.52
##
## The root mean square of the residuals (RMSR) is  0.05
## The df corrected root mean square of the residuals is  0.07
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                 MR1  MR2
## Correlation of (regression) scores with factors   0.98 0.97
## Multiple R square of scores with factors          0.96 0.94
## Minimum correlation of possible factor scores     0.93 0.88
```

```
factores$communalities
```

```
##        V1        V2        V3        V4        V5        V6        V7        V8
## 0.7406732 0.8476869 0.6629813 0.8807983 0.7245791 0.8616111 0.2941650 0.7246741
##        V9       V10
## 0.7626000 0.7502879
```

Como bien hemos dicho anteriormente, las comunalidades son la información que retienen las variables y deben ser superiores a 0.6, por lo que deberíamos eliminar la variable 'V7' con una comunalidad de 0.2941650.

3.2. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

En el Análisis Factorial primero vemos si las variables son normales o no, aceptamos normalidad cuando la significación es mayor que 0.05, si todas las variables son normales debemos probar a utilizar la extracción de máxima verosimilitud:

Analizar - Pruebas no paramétricas - cuadros de dialogo antiguos - Ks de una muestra - Pasamos todas las variables - Aceptar

Como observamos solo existen 2 variables que son normales (prestaciones, V8 y aerodinámic, V10), no utilizaremos la extracción de Máxima Verosimilitud, ya que no nos creemos la normalidad.

Prueba de Kolmogorov-Smirnov para una muestra

			v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
N			20	20	20	20	20	20	20	20	20	20
Parámetros normales[a,b]	Media		3,70	3,40	3,50	2,80	3,70	3,70	3,65	2,85	2,80	2,65
	Desv. estándar		1,342	1,635	1,192	1,576	,923	1,174	1,268	1,387	1,473	1,348
Máximas diferencias extremas	Absoluta		,238	,236	,263	,227	,327	,251	,309	,180	,207	,185
	Positivo		,166	,164	,146	,223	,223	,176	,144	,180	,207	,185
	Negativo		-,238	-,236	-,263	-,227	-,327	-,251	-,309	-,146	-,192	-,142
Estadístico de prueba			,238	,236	,263	,227	,327	,251	,309	,180	,207	,185
Sig. asin. (bilateral)[c]			,004	,005	<,001	,008	<,001	,002	<,001	,089	,025	,071
Sig. Monte Carlo (bilateral)[d]	Sig.		,004	,004	,001	,008	<,001	,002	<,001	,083	,024	,069
	Intervalo de confianza al 99%	Límite inferior	,002	,003	,000	,006	,000	,001	,000	,076	,020	,063
		Límite superior	,005	,006	,002	,010	,000	,003	,001	,090	,028	,076

a. La distribución de prueba es normal.
b. Se calcula a partir de datos.
c. Corrección de significación de Lilliefors.
d. El método de Lilliefors basado en las muestras 10000 Monte Carlo con la semilla de inicio 2000000.

Figure 8: "PRUEBA DE NORMALIDAD"

Analizar - Reducción de dimensiones - Análisis Factorial - Pasamos todas las variables cuantitativas, en este caso todas - En descriptivos seleccionamos KMO y prueba de esfericidad de Barlett, Matriz de covarianzas - En extracción debemos ir probando todos, excepto con el de componentes principales del tema anterior y como hemos dicho, el de máxima verosimilitud, (nos quedamos con el que nos de un % de varianza mayor) - Aceptar

Finalmente, tras probar todas las extracciones: cuadrados mín. no ponderados, min. cuadrados generalizados, factorización ejes principales, factorización alfa y factorización imágenes, nos quedamos con la "Factorización alfa", ya que obtenemos un % acumulado mayor (72.7), <nota: nos fijamos en la última columna.>

Varianza total explicada

	Autovalores iniciales			Sumas de cargas al cuadrado de la extracción		
Factor	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	5,701	57,011	57,011	5,453	54,530	54,530
2	2,069	20,692	77,703	1,817	18,170	72,700
3	,720	7,205	84,908			
4	,548	5,478	90,386			
5	,316	3,158	93,544			
6	,271	2,707	96,251			
7	,146	1,464	97,715			
8	,128	1,280	98,995			
9	,068	,684	99,679			
10	,032	,321	100,000			

Método de extracción: factorización alfa

Figure 9: "RESUMEN ANÁLISIS FACTORIAL"

Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,700
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	163,466
	gl	45
	Sig.	<,001

Figure 10: "PRUEBA DE KMO Y BARLETT"

En esta salida de SPSS podemos ver que el p-value<0.001 no coincide con el valor que nos daba la salida de R, p-value=0.489, según el valor de SPSS, rechazamos la hipótesis nula, lo que quiere decir que las variables si están correlacionadas, la matriz de datos si es válida para continuar con el proceso, mientras que en R no era fiable continuar realizando un análisis factorial. En cambio vemos que el valor de KMO (Kaiser-Meyer-Olkin) es 0,7. En SPSS vemos que tanto el KMO como el Test de Barlett nos indica que continuemos con el análisis factorial.

Matriz de correlaciones

		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
Correlación	v1	1,000	,873	,823	,816	-,501	-,194	,213	-,648	-,645	-,497
	v2	,873	1,000	,729	,829	-,439	-,071	,249	-,784	-,752	-,697
	v3	,823	,729	1,000	,812	-,478	-,226	,192	-,557	-,630	-,540
	v4	,816	,829	,812	1,000	-,550	-,262	,174	-,737	-,789	-,654
	v5	-,501	-,439	-,478	-,550	1,000	,738	,175	,292	,341	,123
	v6	-,194	-,071	-,226	-,262	,738	1,000	,421	,132	,055	-,236
	v7	,213	,249	,192	,174	,175	,421	1,000	-,301	-,180	-,414
	v8	-,648	-,784	-,557	-,737	,292	,132	-,301	1,000	,886	,730
	v9	-,645	-,752	-,630	-,789	,341	,055	-,180	,886	1,000	,785
	v10	-,497	-,697	-,540	-,654	,123	-,236	-,414	,730	,785	1,000

Figure 11: "MATRIZ DE CORRELACIONES"

Matriz de covarianzas

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
v1	1,800	1,916	1,316	1,726	-,621	-,305	,363	-1,205	-1,274	-,900
v2	1,916	2,674	1,421	2,137	-,663	-,137	,516	-1,779	-1,811	-1,537
v3	1,316	1,421	1,421	1,526	-,526	-,316	,289	-,921	-1,105	-,868
v4	1,726	2,137	1,526	2,484	-,800	-,484	,347	-1,611	-1,832	-1,389
v5	-,621	-,663	-,526	-,800	,853	,800	,205	,374	,463	,153
v6	-,305	-,137	-,316	-,484	,800	1,379	,626	,216	,095	-,374
v7	,363	,516	,289	,347	,205	,626	1,608	-,529	-,337	-,708
v8	-1,205	-1,779	-,921	-1,611	,374	,216	-,529	1,924	1,811	1,366
v9	-1,274	-1,811	-1,105	-1,832	,463	,095	-,337	1,811	2,168	1,558
v10	-,900	-1,537	-,868	-1,389	,153	-,374	-,708	1,366	1,558	1,818

Figure 12: "MATRIZ DE VARIANZAS Y COVARIANZAS"

Comunalidades

	Inicial	Extracción
v1	,875	,746
v2	,905	,847
v3	,781	,662
v4	,848	,885
v5	,800	,692
v6	,845	,922
v7	,534	,293
v8	,918	,727
v9	,907	,736
v10	,794	,759

Método de extracción: factorización alfa

Si nos fijamos en ambas salidas, tanto en la que hemos visto antes de R como en la de SPSS, observamos que las comunalidades son parecidas, algunas coinciden, en cambio otras son diferentes. Por ejemplo, la variable 'V5', tiene una comunalidad en R de 0.7245791, mientras que en SPSS tiene 0.692. La variable capacidad 'V7' tanto en R como en SPSS llegamos a la conclusión de que deberíamos eliminarla ya que la extracción nos proporciona un valor de 0.294.

4. ANÁLISIS DE CORRESPONDENCIAS.

El análisis de correspondencias (AC) es una técnica estadística utilizada para analizar la relación entre dos o más variables categóricas. Es una extensión del análisis de componentes principales (CP) que se aplica específicamente a datos de tablas de contingencia, donde se cruzan dos o más variables categóricas.

Asimismo, el análisis de correspondencias es un método de extracción de variables ficticias cuantitativas a partir de cualitativas originales que permite continuar con otras técnicas de análisis de datos con variables cuantitativas. Cuando el número de variables cualitativas es superior a dos se dice Análisis de Correspondencias Múltiple.

En resumen, el análisis de correspondencias simples es una técnica descriptiva que permite resumir una gran cantidad de datos en un número menor de variables incorrelaciones, con la intención de brindar la menor pérdida de información posible. Esta técnica en otras palabras, se utiliza para visualizar gráficamente puntos de fila y puntos de columna en un espacio de baja dimensión. El análisis de correspondencias simples se utiliza comúnmente en la representación de datos que se pueden presentar en forma de tablas de contingencia de dos variables nominales u ordinales.

En este apartado veremmos tres tipos de ejercicios, el primero será un ejercicio de análisis de correspondencias simples en el que nosotros debamos introducir en R los datos y crear nosotros una tabla con los datos, el segundo será también análisis de correspondencias simples pero contando ya con una base de datos, y por último, análisis de correspondencias múltiple.

4.1. Ejercicio 3. Consideramos la tabla que muestra la distribución hipotética de los asientos del parlamento europeo entre los partidos políticos de 5 naciones, realiza un análisis de correspondencias para discutir la relación entre país y partido político. En este ejercicio debemos introducir los datos y ponderar cada caso por su precuencia.

	Dem.crist.	Socialista	Otros
Bélgica	8	9	7
Alemania	39	30	6
Italia	25	11	39
Luxemburgo	3	2	1
Holanda	13	10	2

Figure 13: "TABLA DE DATOS"

Introducimos los datos del ejercicio:

```
datos.acs <- matrix(c(8,9,7,
                      39,30,6,
                      25,11,39,
                      3,2,1,
                      13,10,2),nrow=5,ncol=3,byrow=T)
datos.acs
```

```
##      [,1] [,2] [,3]
## [1,]    8    9    7
## [2,]   39   30    6
## [3,]   25   11   39
## [4,]    3    2    1
## [5,]   13   10    2
```

Añadimos los nombres a la tabla:

```
dimnames(datos.acs)<- list(naciones=c("Bélgica","Alemania","Italia","Luxemburgo","Holanda"),
                           partidos_politicos=c("Dem.crist","Socialista","Otros"))
datos.acs
```

```
##             partidos_politicos
## naciones     Dem.crist Socialista Otros
##   Bélgica            8          9     7
##   Alemania          39         30     6
##   Italia            25         11    39
##   Luxemburgo         3          2     1
##   Holanda           13         10     2
```

```
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 4.2.3
```

```
CrossTable(datos.acs,
           prop.t=F,          # Frecuencia Relativa
           prop.r=F,          # Perfil Fila
           prop.c=F,          # Perfil Columna
           prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  205
##
##
##              | partidos_politicos
##     naciones |  Dem.crist | Socialista |     Otros |  Row Total |
## -------------|------------|------------|-----------|------------|
##      Bélgica |          8 |          9 |         7 |         24 |
## -------------|------------|------------|-----------|------------|
##     Alemania |         39 |         30 |         6 |         75 |
## -------------|------------|------------|-----------|------------|
##       Italia |         25 |         11 |        39 |         75 |
## -------------|------------|------------|-----------|------------|
##   Luxemburgo |          3 |          2 |         1 |          6 |
## -------------|------------|------------|-----------|------------|
##      Holanda |         13 |         10 |         2 |         25 |
## -------------|------------|------------|-----------|------------|
## Column Total |         88 |         62 |         55 |        205 |
## -------------|------------|------------|-----------|------------|
##
##
```

Tabla de contingencias con el paquete gmodels y función CrossTable(). Observamos que tengamos los 205 elementos.

Prueba de Independencia Chi-Cuadrado, si el pvalor<0.05 rechazo la hipótesis nula. Contrastamos la hipótesis nula de independencia entre las dos variables que conforman la tabla de contingencia.

Ho: la matriz de correspondencia no tiene tendencia, es decir, f es igual a f(gorro)

H1: la matriz de correspondencia si tiene tendencia

```
prueba <- chisq.test(datos.acs)
```

```
## Warning in chisq.test(datos.acs): Chi-squared approximation may be incorrect
```

```
prueba
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos.acs
## X-squared = 44.917, df = 8, p-value = 3.816e-07
```

Observamos que obtenemos un p-value= 3.816e-07<0.05, por lo tanto, rechazamos la hipótesis nula, es decir, rechazamos que la matriz de correspondencia no tiene tendencia, por lo tanto, si tiene tendencia.

Tabla de perfiles fila y perfiles columnas

```
# Frecuencia Relativa (fij)
prop.table(datos.acs)
```

```
##             partidos_politicos
## naciones      Dem.crist  Socialista       Otros
##   Bélgica    0.03902439 0.043902439 0.034146341
##   Alemania   0.19024390 0.146341463 0.029268293
##   Italia     0.12195122 0.053658537 0.190243902
##   Luxemburgo 0.01463415 0.009756098 0.004878049
##   Holanda    0.06341463 0.048780488 0.009756098
```

```
# Perfiles Fila
perfiles_fila=prop.table(datos.acs, 1);perfiles_fila
```

```
##             partidos_politicos
## naciones     Dem.crist Socialista     Otros
##   Bélgica    0.3333333  0.3750000 0.2916667
##   Alemania   0.5200000  0.4000000 0.0800000
##   Italia     0.3333333  0.1466667 0.5200000
##   Luxemburgo 0.5000000  0.3333333 0.1666667
##   Holanda    0.5200000  0.4000000 0.0800000
```

```
# Perfiles Columna
perfiles_columna=prop.table(datos.acs, 2);perfiles_columna
```

```
##             partidos_politicos
## naciones      Dem.crist Socialista      Otros
##   Bélgica    0.09090909 0.14516129 0.12727273
##   Alemania   0.44318182 0.48387097 0.10909091
##   Italia     0.28409091 0.17741935 0.70909091
##   Luxemburgo 0.03409091 0.03225806 0.01818182
##   Holanda    0.14772727 0.16129032 0.03636364
```

Para realizar en R el Análisis de Correspondencias Simples (ACS) debemos instalar las siguientes librerías.

```
#install.packages('ade4')
library(ade4)
```

```
## Warning: package 'ade4' was built under R version 4.3.3
```

```
#install.packages('factoextra')
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.3

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
#install.packages('FactoMineR')
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.3

##
## Attaching package: 'FactoMineR'

## The following object is masked from 'package:ade4':
##
##     reconst
```

Para realizar un análisis de correspondencias simples se hace uso de la función CA perteneciente al paquete FactomineR.

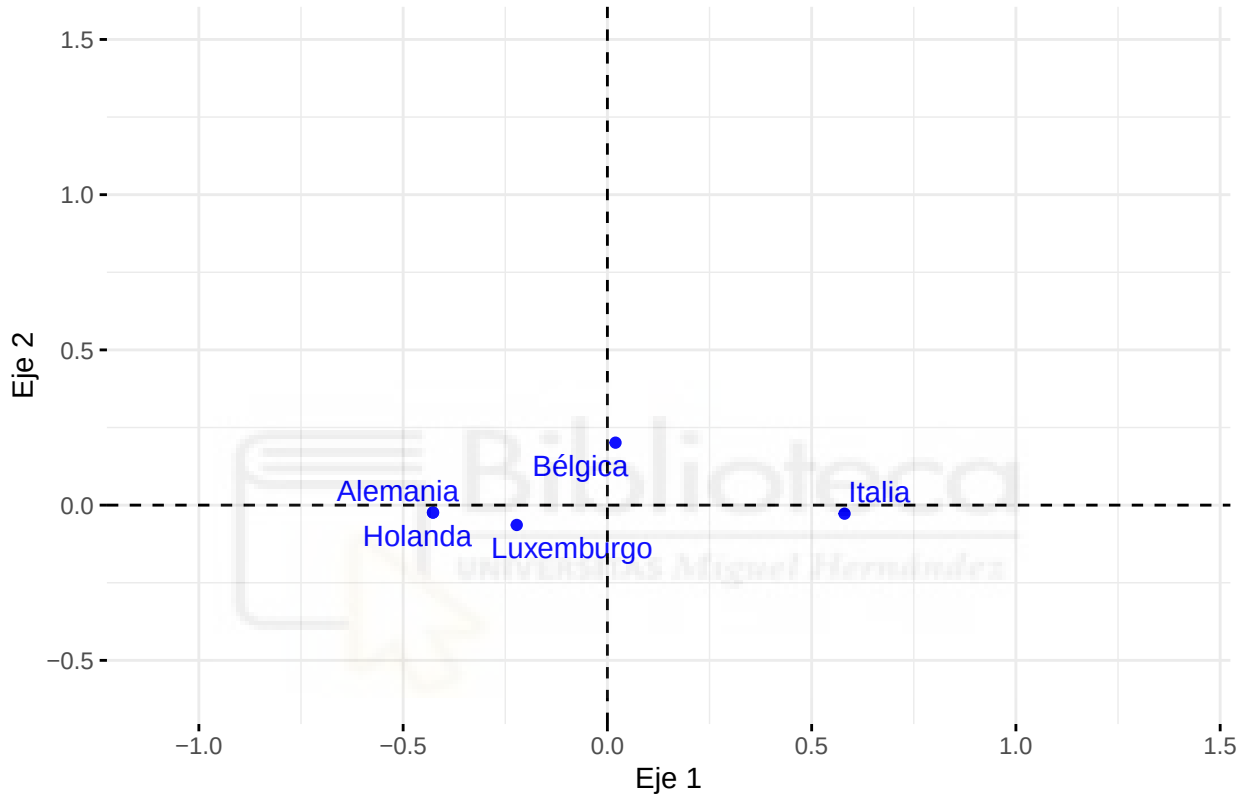
```
#Análisis de correspondencia simple
ACS <- CA(datos.acs, graph = FALSE)
#% de varianza explicado
valores_propios=ACS$eig; valores_propios
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.213694026             97.529295                           97.52929
## dim 2 0.005413501              2.470705                          100.00000
```

Con la intención de identificar el número de componentes a utilizar se realiza el estudio del porcentaje de varianza explicado por ejes, es importante mencionar que el número de dimensiones está asociado con la cantidad de columnas, ya que por lo general siempre es menor que el número de filas; teniendo en cuenta que dentro del análisis una de las columnas termina siendo combinación lineal de las demás se obtendrán en total p-1 dimensiones y en este caso, son 2 ya que el número total de columnas es 3.

Análisis puntos fila

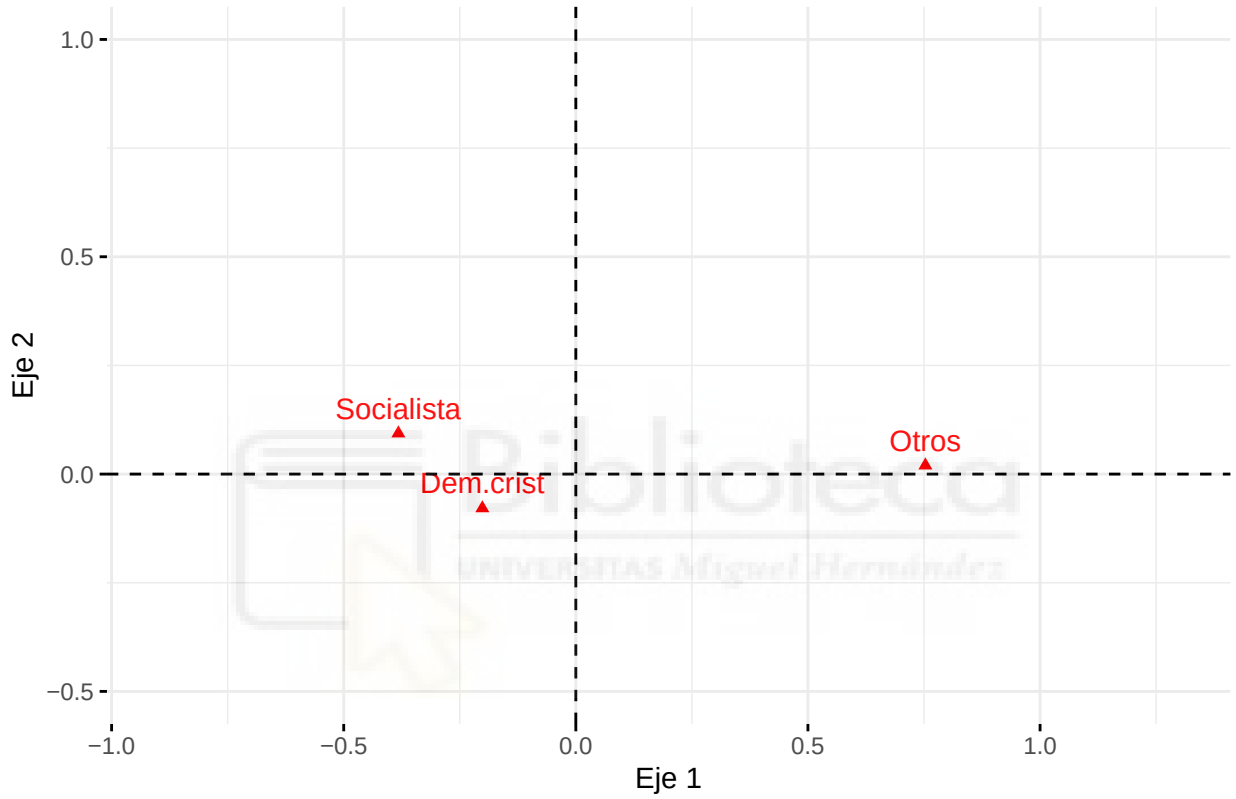
```
#Perfiles fila
variables_fila=get_ca_row(ACS)
#Nube de individuos fila
fviz_ca_row(ACS, repel = TRUE)+ggtitle("") + ylab("Eje 2")+xlab("Eje 1")+ylim(-0.6,1.5)+xlim(-1.1,1.4)
```



A partir de la nube de puntos fila se observa que las categorías de las 5 naciones (Bélgica, 1, Alemania, 2, Italia, 3, Luxemburgo, 4, Holanda, 5) guardan ciertas distancias entre si y resulta lógico, ya que en términos generales y de acuerdo con estudios hay ciertas similitudes pero entre 2 y 5, es decir, que Alemania y Holanda comparten ciertas características pero a su vez difieren de las demás naciones; con respecto a la contribución que pueden presentar los puntos fila a la construcción de los ejes factoriales se observa que posiblemente las categorías 2,3,4,5 (naciones Alemania, Italia, Luxemburgo y Holanda) podrían estar aportando mayor información a este eje y para el segundo se podría pensar que la categoría 1 (nación Bélgica), ya que se encuentra más en la dirección de este eje. No obstante, se comprobará si esta afirmación es cierta en la sección de contribuciones.

Análisis puntos columna

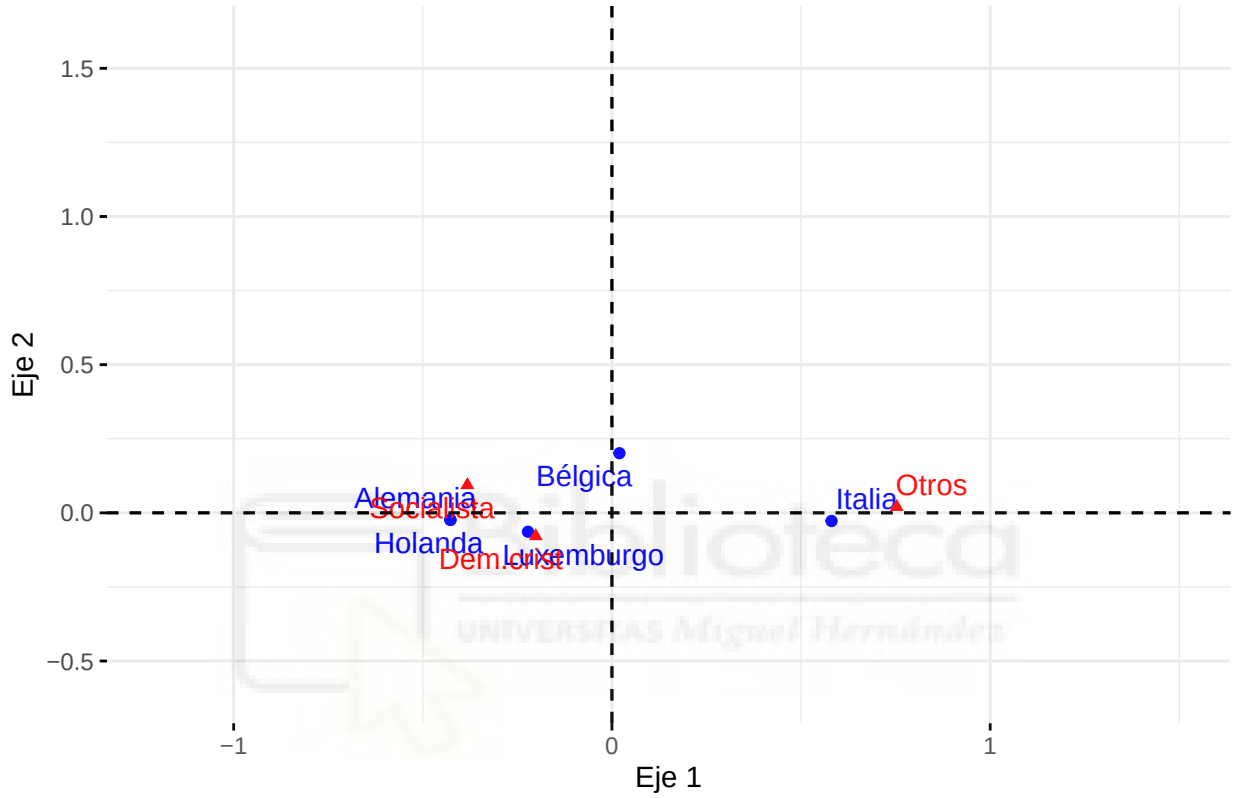
```
#Perfiles columna
variables_columna=get_ca_col(ACS)
#Nube de individuos columna
fviz_ca_col(ACS)+ggtitle("")+ylab("Eje 2")+xlab("Eje 1")+ylim(-0.5,1)+xlim(-0.9,1.3)
```



De la nube de puntos columna se observa que los asientos del parlamento europeo entre los partidos políticos (Dem.crist., V1, Socialista, V2, Otros, V3) se encuentran un poco distantes entre si, aunque V1 Y V2 parecen estar más juntos y tiene sentido, ya que si una persona dice ser de otro partido político no es posible relacionarla con otra categoría pues socialista y otros se podrían tomar como deterministas y el partido Dem.crist indica la posición intermedia de las personas; con respecto a la contribución que pueden presentar los puntos columna a la construcción de los ejes factoriales se observa que posiblemente la categoría V1 podría estar aportando mayor información a este eje y para el segundo se podría pensar que las categorías V2 y V3, ya que se encuentra más en la dirección de este eje.

Representación simultánea

```
#Representación simultánea
fviz_ca_biplot(ACS, repel = TRUE)+ggtitle("")+ylab("Eje 2")+xlab("Eje 1")+ylim(-0.6,1.6)+xlim(-1.2,1.5)
```



De la representación simultánea se puede observar que cuanto más lejos del origen de coordenadas, más influyentes son, por ejemplo, los más influyentes en este caso son Italia y Otros. El menos influyente (más cerca del origen) podríamos decir que es Bélgica y Dem.crist. Además también podemos decir que existe una relación entre Dem.crist. y Luxemburgo. En síntesis, es posible concluir que efectivamente las variables estudiadas no presentan independencia entre si y por tanto, es posible discriminar la distribución hipotética de lso asientos del parlamento europeo entre los partidos políticos de las 5 naciones.

Contribuciones:

```
#Contribuciones por fila
contribuciones_fila=variables_fila$contrib;contribuciones_fila
```

```
##                  Dim 1     Dim 2
## Bélgica     0.02161882 87.565468
## Alemania   31.19826158  3.837176
## Italia     57.70639161  5.116426
## Luxemburgo  0.67430746  2.201871
## Holanda    10.39942053  1.279059
```

```
#contribuciones por columna
contribuciones_columna=variables_columna$contrib; contribuciones_columna
```

```
##                Dim 1    Dim 2
## Dem.crist   8.142976 48.93020
## Socialista 20.673523 49.08257
## Otros      71.183501  1.98723
```

Para los puntos fila se observa que las naciones correspondientes a Alemania, Italia y Holanda son las que contribuyen más a la construcción del primer eje y para el segundo eje, se observa que Bélgica es la que más contribuye a su construcción con un valor de 87.57%; con respecto a la calidad de representación se observa que los estratos Alemania, Italia y Holanda son los que se encuentran mejor representados en el primer eje y Bélgica es el que representa en mejor medida la dimensión 2.

Para los puntos columna se observa que la categoría V3 (Otros) es la que más contribuye a la formación del primer eje, mientras que V1 y V2 (Dem.crist y Socialista, respectivamente) contribuyen a la construcción del segundo eje.

4.2. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Datos - Ponderar casos - Pasamos las variables - Aceptar

Analizar - Reducción de dimensiones - Análisis de correspondencias - En actividades seleccionamos perfiles fila y columna - Aceptar

Tabla de correspondencias

País	Partido			
	Demócrata Cristiano	Socialista	Otros	Margen activo
Bélgica	8	9	7	24
Alemania	39	30	6	75
Italia	25	11	39	75
Luxemburgo	3	2	1	6
Holanda	13	10	2	25
Margen activo	88	62	55	205

Figure 14: "TABLA DE CONTINGENCIAS"

Resumen

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulativa	Desviación estándar	Correlación 2
1	,462	,214			,975	,975	,060	-,073
2	,074	,005			,025	1,000	,069	
Total		,219	44,917	<,001[a]	1,000	1,000		

a. 8 grados de libertad

Figure 15: "PRUEBA DE INDEPENDENCIA"

Lo primero es mirar el pvalor, y como es menor que 0.05 entonces rechazamos la hipótesis nula de que f=fgorro, por lo tanto, sí hay tendencia. En "proporción de inercia contabilizado para", vemos que la primera dimensión es la que más tendencia tiene (0.975), está muy desproporcionada, el segundo eje casi no tiene importanci (0.025). La nube d puntos será bastante horizontal, porque se describe más en el primer eje.

Perfiles de fila

País	Partido: Demócrata Cristiano	Socialista	Otros	Margen activo
Bélgica	,333	,375	,292	1,000
Alemania	,520	,400	,080	1,000
Italia	,333	,147	,520	1,000
Luxemburgo	,500	,333	,167	1,000
Holanda	,520	,400	,080	1,000
Masa	,429	,302	,268	

Perfiles de columna

País	Partido: Demócrata Cristiano	Socialista	Otros	Masa
Bélgica	,091	,145	,127	,117
Alemania	,443	,484	,109	,366
Italia	,284	,177	,709	,366
Luxemburgo	,034	,032	,018	,029
Holanda	,148	,161	,036	,122
Margen activo	1,000	1,000	1,000	

Figure 16: "PERFILES DE FILA Y COLUMNA"

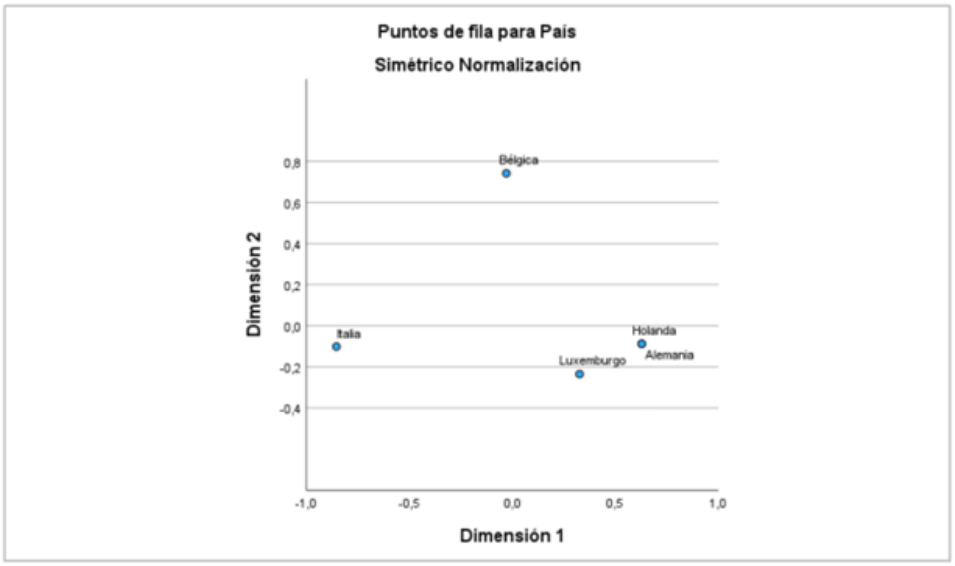


Figure 17: "GRÁFICO DE FILA"

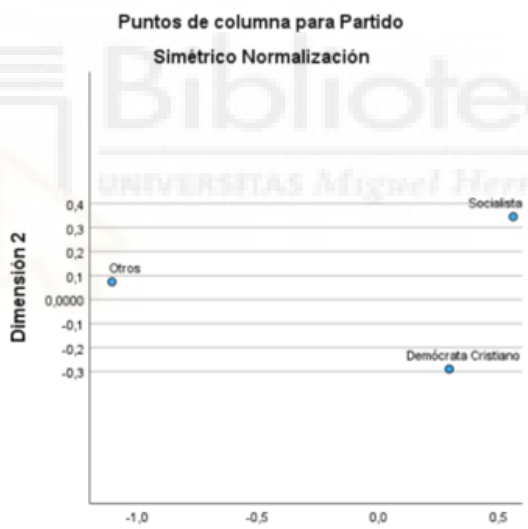


Figure 18: "GRÁFICO DE COLUMNA"

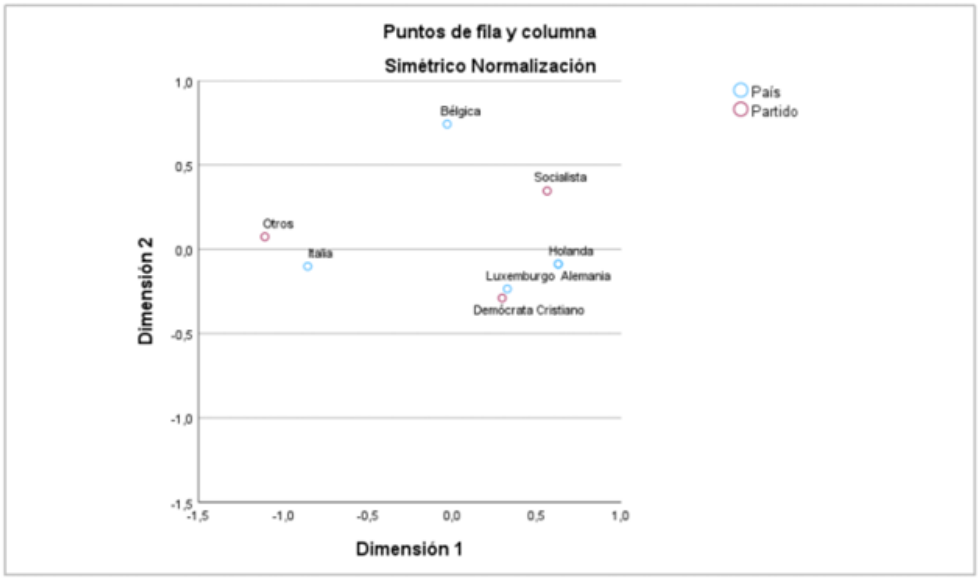


Figure 19: "GRÁFICO PUNTOS DE FILA Y COLUMNA"

Cuanto más lejos del origen de coordenadas, más influyentes son: Italia y Otros son los dos más influyentes y la menos influyente Luxemburgo.

Puntos de fila generales[a]

País	Masa	Puntuación en dimensión 1	Puntuación en dimensión 2	Inercia	Contribución: Del punto en la inercia de dimensión 1	Contribución: Del punto en la inercia de dimensión 2	Contribución: De la dimensión en la inercia del punto 1	Contribución: De la dimensión en la inercia del punto 2	Contribución: De la dimensión en la inercia del punto Total
Bélgica	,117	-,029	,742	,005	,000	,876	,010	,990	1,000
Alemania	,366	,628	-,088	,067	,312	,038	,997	,003	1,000
Italia	,366	-,854	-,101	,124	,577	,051	,998	,002	1,000
Luxemburgo	,029	,326	-,235	,002	,007	,022	,924	,076	1,000
Holanda	,122	,628	-,088	,022	,104	,013	,997	,003	1,000
Total activo	1,000			,219	1,000	1,000			

a. Normalización simétrica

Puntos de columna generales[a]

Partido	Masa	Puntuación en dimensión 1	Puntuación en dimensión 2	Inercia	Contribución: Del punto en la inercia de dimensión 1	Contribución: Del punto en la inercia de dimensión 2	Contribución: De la dimensión en la inercia del punto 1	Contribución: De la dimensión en la inercia del punto 2	Contribución: De la dimensión en la inercia del punto Total
Demócrata Cristiano	,429	,296	-,290	,020	,081	,489	,868	,132	1,000
Socialista	,302	,562	,346	,047	,207	,491	,943	,057	1,000
Otros	,268	-1,107	,074	,152	,712	,020	,999	,001	1,000
Total activo	1,000			,219	1,000	1,000			

a. Normalización simétrica

Figure 20: "PUNTOS GENERALES"

Conclusiones:

Puntos de fila generales: Si queremos ver en que país hay más encuestados debemos mirar las masas, en este caso son Alemania e Italia con un valor de 0'366.

Si queremos saber que país ayuda más a definir la primera dimensión nos fijamos en la contribución del punto en la inercia de la dimensión 1, en este caso vemos que la que más influye es Italia con un valor de 0'577.

Puntos de columna generales: Si queremos saber quien es el partido que ayuda a definir más la primera dimensión, es decir, la que más influye, nos fijamos en la contribución del punto en la inercia de la dimensión 1, y vemos que Otros influye más con un valor de 0'712 y el que menos influye es Demócrata Cristiano con un valor de 0'081.

En la tabla resumen, se observa que la segunda dimensión acumula muy poca inercia (0'005), para decir a quien corresponde se ve en contribución de la dimensión en la inercia del lpunto en la segunda dimensión. Entonces, vemos que solamente Bélgica es quien justifica que exista una segunda dimensión (0'99).

4.3. Ejercicio 4. Realiza un análisis de correspondencias para estudiar la relación entre país de origen y número de cilindros de los datos en coches.sav.

Como el ejercicio nos dice que estudiemos la relación entre dos variables, seguimos con un análisis de correspondencias simple. En este ejercicio no es necesario introducir la tabla, la dificultad podría ser que debemos seleccionar las variables de las cuales queremos estudiar la relación de la base de datos.

Leemos la base de datos:

```
library(foreign)
datos <- read.spss("coches.sav",
                   use.value.labels = T,
                   to.data.frame=TRUE)

summary(datos)
```

```
##     consumo          motor            cv              peso
##  Min.   : 5.00   Min.   :  66   Min.   : 46.00   Min.   : 244.0
##  1st Qu.: 8.00   1st Qu.:1708   1st Qu.: 75.75   1st Qu.: 741.2
##  Median :10.00   Median :2434   Median : 95.00   Median : 936.5
##  Mean   :11.23   Mean   :3180   Mean   :104.83   Mean   : 989.5
##  3rd Qu.:13.00   3rd Qu.:4806   3rd Qu.:129.25   3rd Qu.:1203.8
##  Max.   :26.00   Max.   :7456   Max.   :230.00   Max.   :1713.0
##  NA's   :8                      NA's   :6
##       acel            año            origen          cilindr      destino
##  Min.   : 8.00   Min.   : 0.00   EE.UU.:253   3 cilindros:  4   EEUU  : 85
##  1st Qu.:13.62   1st Qu.:73.00   Europa: 73   4 cilindros:207   OCDE  :180
##  Median :15.50   Median :76.00   Japón : 79   5 cilindros:  3   Japón : 79
##  Mean   :15.50   Mean   :75.75   NA's  :  1   6 cilindros: 84   Europa: 62
##  3rd Qu.:17.07   3rd Qu.:79.00                8 cilindros:107
##  Max.   :24.80   Max.   :82.00                NA's       :  1
##
```

La base de datos cuenta con 9 variables (consumo, motor, cv, peso, acel, año, origen, cilindr y destino). La tres últimas respectivamente son variables cualitativas mientras que las demás son cuantitativas y todas se encuentran en el mismo orden de magnitud, todas ellas cuentan con la cantidad de 405 observaciones.

Observamos la base de datos, creamos una tabla con las variables que queremos estudiar su relación, en este caso, origen y cilindr:

```
attach(datos)
table(origen)
```

```
## origen
## EE.UU. Europa  Japón
##    253     73     79
```

```
table(cilindr)
```

```
## cilindr
## 3 cilindros 4 cilindros 5 cilindros 6 cilindros 8 cilindros
##           4         207           3          84         107
```

```
addmargins(table(origen,cilindr))
```

```
##         cilindr
## origen   3 cilindros 4 cilindros 5 cilindros 6 cilindros 8 cilindros Sum
##   EE.UU.          0         72          0         74         107 253
##   Europa          0         66          3          4           0  73
##   Japón           4         69          0          6           0  79
##   Sum             4        207          3         84         107 405
```

Convertimos los datos en una matrix sino no podremos realizar el análisis de correspondencias simple:

```
datos.acs1 <- as.matrix(table(origen,cilindr))
datos.acs1
```

```
##         cilindr
## origen   3 cilindros 4 cilindros 5 cilindros 6 cilindros 8 cilindros
##   EE.UU.          0         72          0         74         107
##   Europa          0         66          3          4           0
##   Japón           4         69          0          6           0
```

Carga de paquetes

```
#install.packages('ade4')
library(ade4)
#install.packages('factoextra')
library(factoextra)
```

Observamos que tengamos los 405 elementos.

```
# Tabla con el paquete gmodels y función CrossTable()
library(gmodels)
CrossTable(datos.acs1,
           prop.t=F,          # Frecuencia Relativa
           prop.r=F,          # Perfil Fila
           prop.c=F,          #Perfil Columna
           prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  405
##
##
##              | cilindr
##       origen | 3 cilindros | 4 cilindros | 5 cilindros | 6 cilindros | 8 cilindros |   Row Total |
## -------------|-------------|-------------|-------------|-------------|-------------|-------------|
##       EE.UU. |           0 |          72 |           0 |          74 |         107 |         253 |
## -------------|-------------|-------------|-------------|-------------|-------------|-------------|
##       Europa |           0 |          66 |           3 |           4 |           0 |          73 |
## -------------|-------------|-------------|-------------|-------------|-------------|-------------|
##        Japón |           4 |          69 |           0 |           6 |           0 |          79 |
## -------------|-------------|-------------|-------------|-------------|-------------|-------------|
## Column Total |           4 |         207 |           3 |          84 |         107 |         405 |
## -------------|-------------|-------------|-------------|-------------|-------------|-------------|
##
##
```

En primer lugar, resulta necesario identificar si realmente existe relación entre las variables seleccionadas, ya que en caso de no existir no resulta conveniente hacer uso de la técnica de Análisis de Correspondencias Simple (ACS), para comprobar esto se hará uso de la prueba de independencia Ji cuadrado, la cuál se ilustra a continuación. Prueba de independencia:

```
chisq.test(datos.acs1)
```

```
## Warning in chisq.test(datos.acs1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos.acs1
## X-squared = 185.79, df = 8, p-value < 2.2e-16
```

Es posible observar que el estadístico de prueba Ji cuadrado cuyo valor, p-value $<$ 2.2e-16, rechazamos la hipótesis nula, es decir, rechazamos que la matriz de correspondencia no tiene tendencia, por lo tanto, por lo tanto, resulta conveniente hacer uso de la técnica ACS para el conjunto de datos.

Tabla de perfiles fila y perfiles columnas:

```
# Frecuencia Relativa (fij)
prop.table(datos.acs1)
```

```
##         cilindr
## origen   3 cilindros 4 cilindros 5 cilindros 6 cilindros 8 cilindros
##   EE.UU. 0.000000000 0.177777778 0.000000000 0.182716049 0.264197531
##   Europa 0.000000000 0.162962963 0.007407407 0.009876543 0.000000000
##   Japón  0.009876543 0.170370370 0.000000000 0.014814815 0.000000000
```

```
# Perfiles Fila
perfiles_fila=prop.table(datos.acs1, 1);
perfiles_fila
```

```
##         cilindr
## origen   3 cilindros 4 cilindros 5 cilindros 6 cilindros 8 cilindros
##   EE.UU.  0.00000000  0.28458498  0.00000000  0.29249012  0.42292490
##   Europa  0.00000000  0.90410959  0.04109589  0.05479452  0.00000000
##   Japón   0.05063291  0.87341772  0.00000000  0.07594937  0.00000000
```

```
# Perfiles Columna
perfiles_columna=prop.table(datos.acs1, 2);
perfiles_columna
```

```
##         cilindr
## origen   3 cilindros 4 cilindros 5 cilindros 6 cilindros 8 cilindros
##   EE.UU.  0.00000000  0.34782609  0.00000000  0.88095238  1.00000000
##   Europa  0.00000000  0.31884058  1.00000000  0.04761905  0.00000000
##   Japón   1.00000000  0.33333333  0.00000000  0.07142857  0.00000000
```

Como hemos dicho en el ejercicio anterior, para realizar un análisis de correspondencias simple necesitamos hacer uso de la función CA que pertenece al paquete FactomineR:

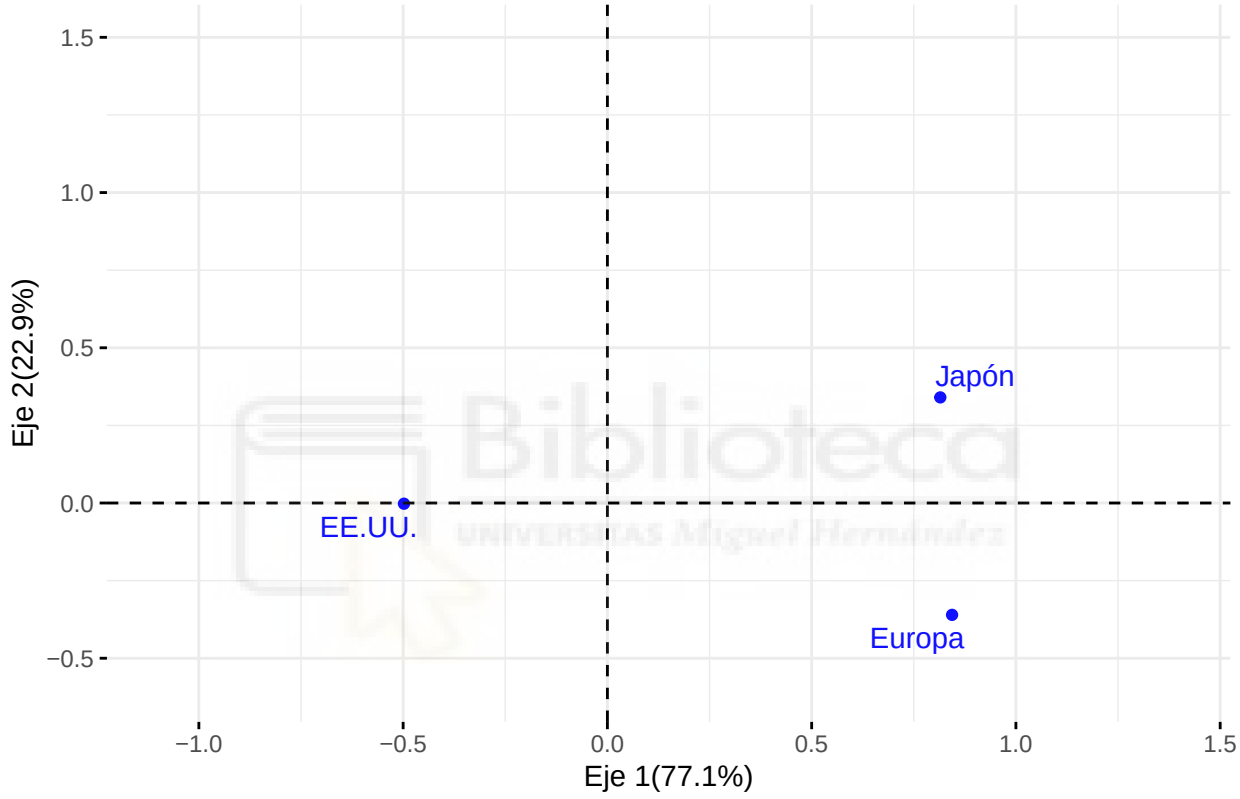
```
#install.packages('FactoMineR')
library(FactoMineR)
#Análisis de correspondencia simple
ACS <- CA(datos.acs1, graph = FALSE)
#% de varianza explicado
valores_propios=ACS$eig; valores_propios
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.41276802              89.97669                           89.97669
## dim 2 0.04598191              10.02331                          100.00000
```

Con la intención de identificar el número de componentes a utilizar se realiza el estudio del porcentaje de varianza explicado por ejes, es importante mencionar que el número de dimensiones está asociado con la cantidad de columnas, ya que por lo general siempre es menor que el número de filas; teniendo en cuenta que dentro del análisis una de las columnas termina siendo combinación lineal de las demás se obtendrán en total p-1 dimensiones y en este caso, son 2 ya que el número total de columnas es 3.

Análisis puntos fila

```
#Perfiles fila
variables_fila=get_ca_row(ACS)

#Nube de individuos fila
fviz_ca_row(ACS, repel = TRUE)+ggtitle("") + ylab("Eje 2(22.9%)")+xlab("Eje 1(77.1%)")+ylim(-0.6,1.5)+x
```



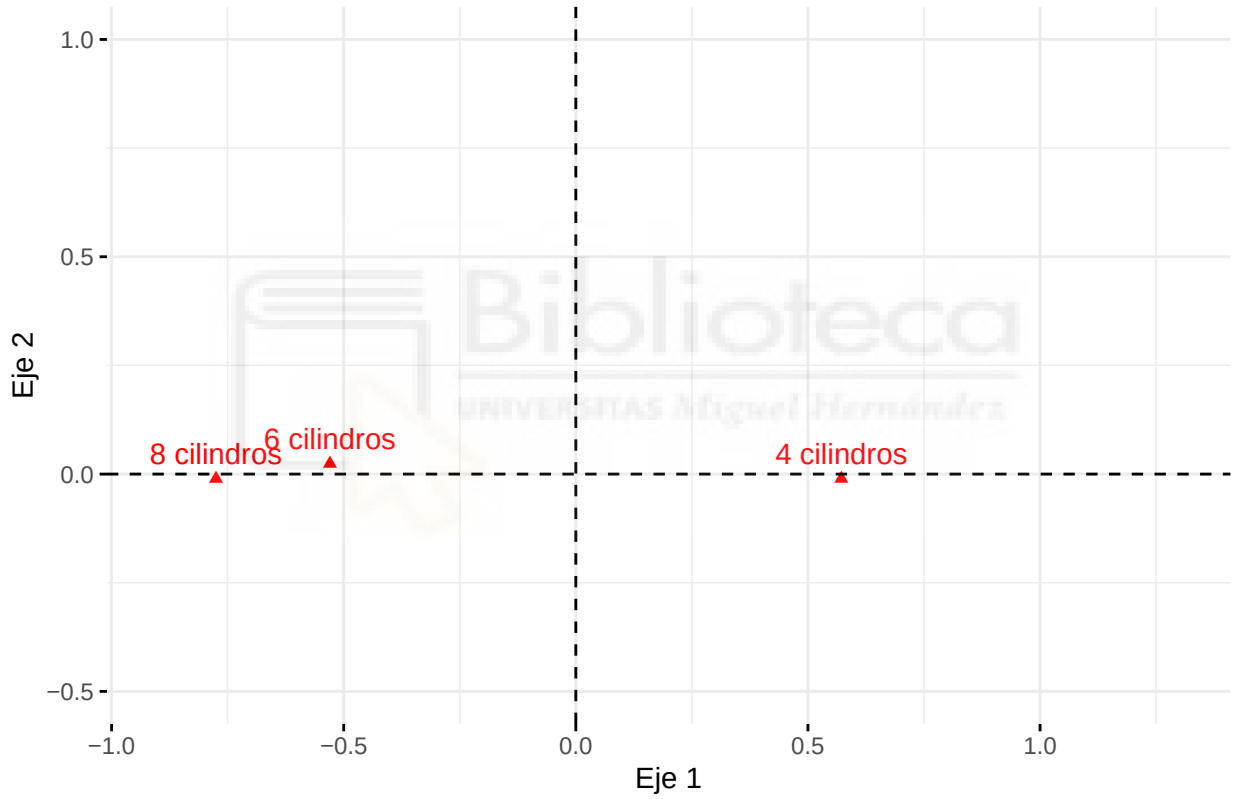
A partir de la nube de puntos fila se observa que las categorías EE.UU, Japón y Europa guardan ciertas distancias entre si; con respecto a la contribución que pueden presentar los puntos fila a la construcción de los ejes factoriales se observa que posiblemente los paises de origen EE.UU y Europa podrían estar aportando mayor información a este eje y para el segundo se podría pensar que el pais de origen Japón, ya que se encuentra más en la dirección de este eje. No obstante, se comprobará si esta afirmación es cierta en la sección de contribuciones.

Análisis puntos columna

```
#Perfiles columna
variables_columna=get_ca_col(ACS)
#Nube de individuos columna
fviz_ca_col(ACS)+ggtitle("")+ylab("Eje 2")+xlab("Eje 1")+ylim(-0.5,1)+xlim(-0.9,1.3)
```

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_text()‘).
```



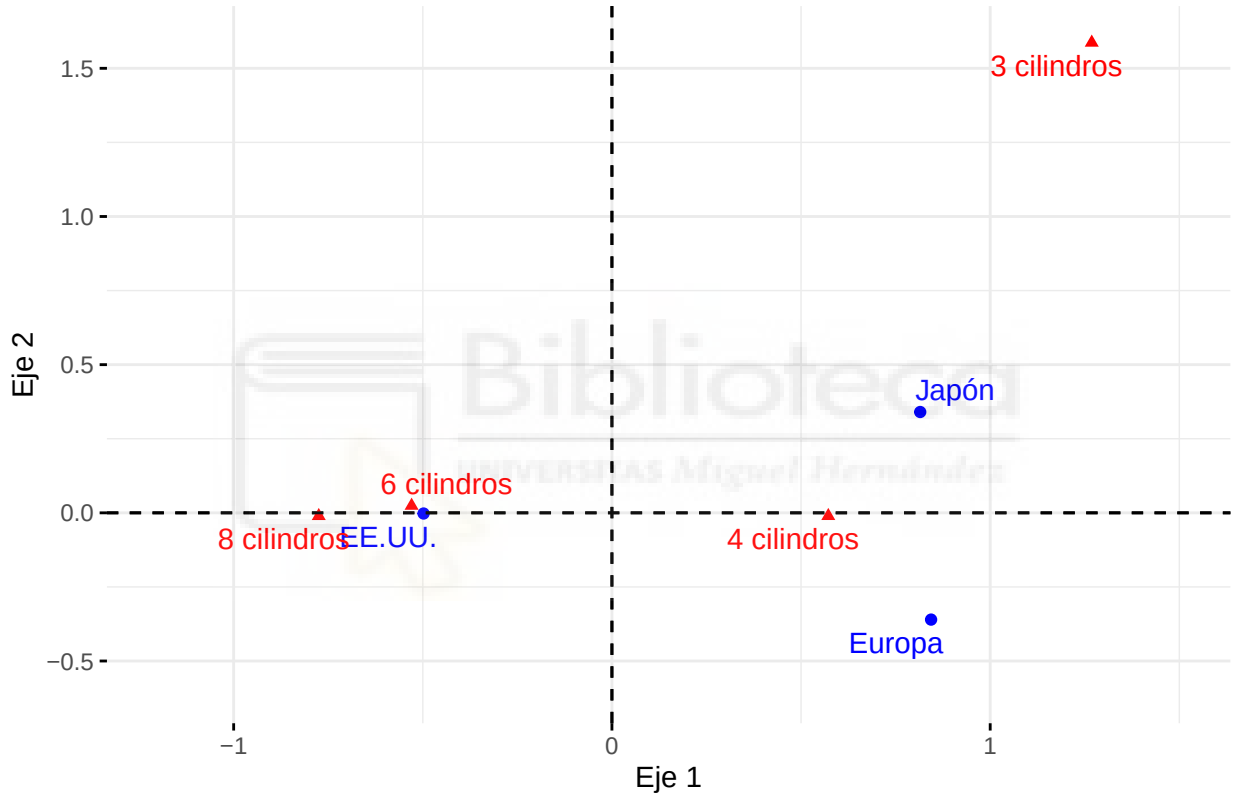
De la nube de puntos columna se observa que las categorías de los cilindros se encuentran un poco distantes entre si y tiene sentido, ya que si un coche tiene 4 cilindros no es posible relacionarla con otra categoría pues 8 y 4 cilindros se podrían tomar como deterministas y 6 cilindros indica la posición intermedia de los cilindros que tiene un coche; con respecto a la contribución que pueden presentar los puntos columna a la construcción de los ejes factoriales se observa que posiblemente las categorías 8 y 4 podrían estar aportando mayor información a este eje y para el segundo se podría pensar que la categoría 6, ya que se encuentra más en la dirección de este eje.

Representación simultánea

```
#Representación simultánea
fviz_ca_biplot(ACS, repel = TRUE)+ggtitle("")+ylab("Eje 2")+xlab("Eje 1")+ylim(-0.6,1.6)+xlim(-1.2,1.5)
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 1 rows containing missing values (‘geom_text_repel()‘).
```



De la representación simultánea se puede observar que cuanto más lejos del origen de coordenadas, más influyentes son, por ejemplo, los más influyentes en este caso son EE.UU y 4 cilindros. EE.UU, Europa y Japón intervienen, es decir, influyen más o menos igual en la primera dimensión, mientras que en la segunda dimensión, solo tenemos Europa y Japón. En síntesis, es posible concluir que efectivamente las variables estudiadas no presentan independencia entre si y por tanto, es posible discriminar la distribución hipotética de lso asientos del parlamento europeo entre los partidos políticos de las 5 naciones.

Contribuciones:

```
#Contribuciones por fila
contribuciones_fila=variables_fila$contrib;contribuciones_fila
```

```
##          Dim 1        Dim 2
## EE.UU. 37.5238  0.007066021
## Europa 31.0893 50.886003786
## Japón  31.3869 49.106930193
```

```
#contribuciones por columna
contribuciones_columna=variables_columna$contrib; contribuciones_columna
```

```
##                 Dim 1      Dim 2
## 3 cilindros  3.850129 54.0740245
## 4 cilindros 40.512911  0.1232967
## 5 cilindros  3.095304 45.4788787
## 6 cilindros 14.094520  0.2588094
## 8 cilindros 38.447137  0.0649907
```

Para los puntos fila se observa que EE.UU es la que contribuye más a la construcción del primer eje y para el segundo eje, se observa que Europa y Japón son las que más contribuyen a su construcción con un valor de 50.88% y 49.11% respectivamente; con respecto a la calidad de representación se observa que EE.UU es el que que se encuentra mejor representado en el primer eje y Europa y Japón son los que representan en mejor medida la dimensión 2.

Para los puntos columna se observa que 4, 6 y 8 cilindros respectivamente son los que más contribuyen a la formación del primer eje, mientras que 3 y 5 cilindros contribuyen a la construcción del segundo eje.

4.4. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Datos - Ponderar casos - Pasamos las variables - Aceptar

Analizar - Reducción de dimensiones - Análisis de correspondencias - Definimos el rango, primero debemos mirar las etiquetas de los países y cilindros, vemos que en el número de cilindros el rango va desde 3 hasta 8, excepto 7, por lo que debemos marcar el 7 como una categoría suplementaria - En actividades seleccionamos perfiles fila y columna - Aceptar

Tabla de correspondencias

	Número de cilindros						
País de origen	3 cilindros	4 cilindros	5 cilindros	6 cilindros	7[a]	8 cilindros	Margen activo
EE.UU.	0	72	0	74	0	107	253
Europa	0	66	3	4	0	0	73
Japón	4	69	0	6	0	0	79
Margen activo	4	207	3	84		107	405

a. Columna complementaria

Figure 21: "TABLA DE CONTINGENCIAS"

Resumen

					Proporción de inercia		Valor singular de confianza	
Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Contabilizado para	Acumulativa	Desviación estándar	Correlación 2
1	,642	,413			,900	,900	,031	,109
2	,214	,046			,100	1,000	,040	
Total		,459	185,794	<,001[a]	1,000	1,000		

a. 8 grados de libertad

Figure 22: "PRUEBA DE INDEPENDENCIA"

Vemos que el valor de Chi cuadrado es adecuado y la significación también.

Perfiles de fila

País de origen	Número de cilindros						
	3 cilindros	4 cilindros	5 cilindros	6 cilindros	7[a]	8 cilindros	Margen activo
EE.UU.	,000	,285	,000	,292	,000	,423	1,000
Europa	,000	,904	,041	,055	,000	,000	1,000
Japón	,051	,873	,000	,076	,000	,000	1,000
Masa	,010	,511	,007	,207	,000	,264	

a. Columna complementaria

Perfiles de columna

País de origen	Número de cilindros						
	3 cilindros	4 cilindros	5 cilindros	6 cilindros	7[a]	8 cilindros	Masa
EE.UU.	,000	,348	,000	,881	,000	1,000	,625
Europa	,000	,319	1,000	,048	,000	,000	,180
Japón	1,000	,333	,000	,071	,000	,000	,195
Margen activo	1,000	1,000	1,000	1,000	,000	1,000	

a. Columna complementaria

Figure 23: "PERFILES DE FILA Y COLUMNA"

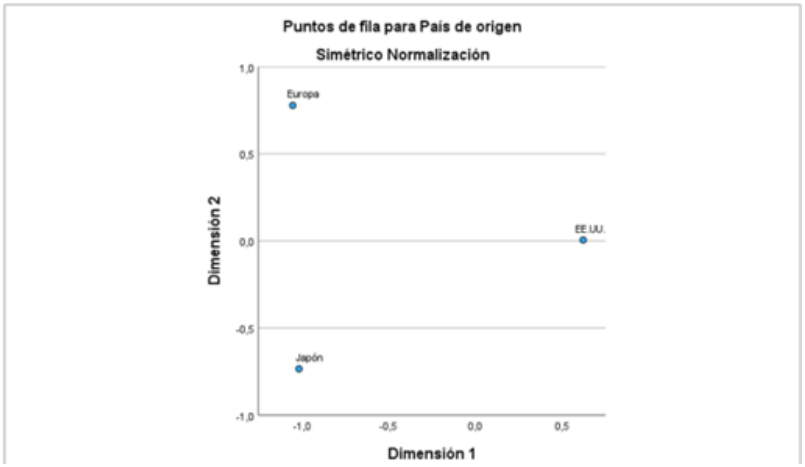


Figure 24: "GRÁFICO DE FILA"

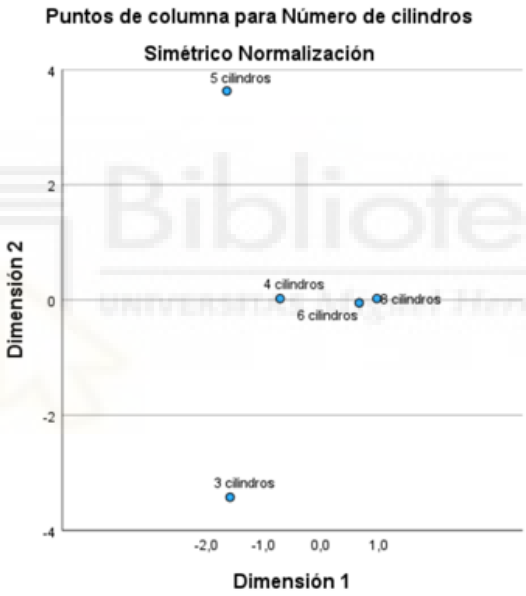


Figure 25: "GRÁFICO DE COLUMNA"

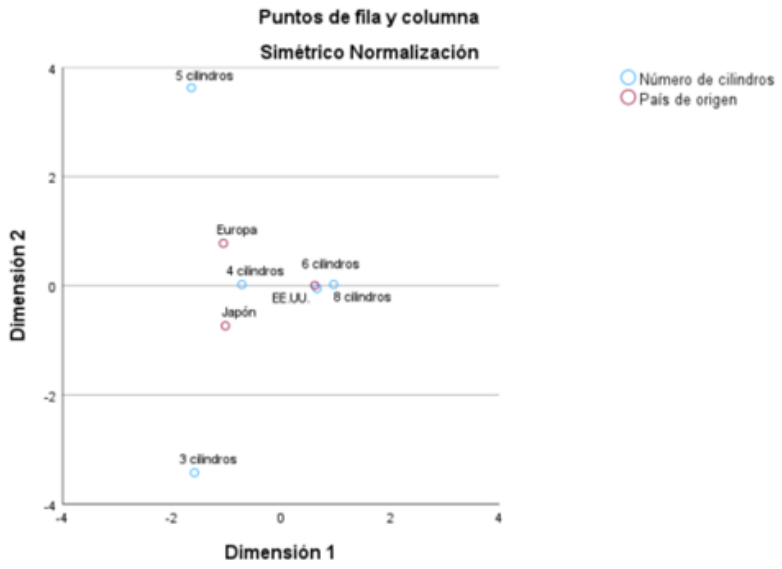


Figure 26: "GRÁFICO SIMULTÁNEO"

En el primer eje, las 2 mas influyentes son: Estados unidos y 4 cilindros, que no está ni arriba ni abajo por ser solo el 10%. Estados unidos, Europa y Japón intervienen (influyen) más o menos igual en la primera dimensión. En la segunda dimensión, solo tenemos Europa y Japón.

Puntos de fila generales[a]

País de origen	Masa	Puntuación en dimensión 1	Puntuación en dimensión 2	Inercia	Contribución: Del punto en la inercia de dimensión 1	Contribución: Del punto en la inercia de dimensión 2	Contribución: De la dimensión en la inercia del punto 1	Contribución: De la dimensión en la inercia del punto 2	Contribución: De la dimensión en la inercia del punto Total
EE.UU.	,625	,621	,005	,155	,375	,000	1,000	,000	1,000
Europa	,180	-1,053	,778	,152	,311	,509	,846	,154	1,000
Japón	,195	-1,017	-,735	,152	,314	,491	,852	,148	1,000
Total activo	1,000			,459	1,000	1,000			

a. Normalización simétrica

Puntos de columna generales[a]

Número de cilindros	Masa	Puntuación en dimensión 1	Puntuación en dimensión 2	Inercia	Contribución: Del punto en la inercia de dimensión 1	Contribución: Del punto en la inercia de dimensión 2	Contribución: De la dimensión en la inercia del punto 1	Contribución: De la dimensión en la inercia del punto 2	Contribución: De la dimensión en la inercia del punto Total
3 cilindros	,010	-1,583	-3,426	,041	,039	,541	,390	,610	1,000
4 cilindros	,511	-,714	,023	,167	,405	,001	1,000	,000	1,000
5 cilindros	,007	-1,638	3,628	,034	,031	,455	,379	,621	1,000
6 cilindros	,207	,661	-,052	,058	,141	,003	,998	,002	1,000
7[b]	,000	.	.	.	,000	,000	.	.	.
8 cilindros	,264	,967	,023	,159	,384	,001	1,000	,000	1,000
Total activo	1,000			,459	1,000	1,000			

a. Normalización simétrica

b. Punto complementario

Figure 27: "PUNTOS DE FILA/COLUMNA GENERALES"

4.5. Ejercicio 5. Realiza un análisis de correspondencias múltiple con el fichero social.sav.

El análisis de correspondencia múltiple (ACM) es una extensión del análisis de correspondencia simple (ACS) para resumir y visualizar una tabla de datos que contiene más de dos variables categóricas. También puede verse como una generalización del análisis de componentes principales (CP) cuando las variables a analizar son categóricas en lugar de cuantitativas.

Leemos la base de datos:

```
library(foreign)
datos <- read.spss("social.sav",
                   use.value.labels = F,
                   to.data.frame=TRUE)
summary(datos)
```

```
##     INTENSID        FRECUENC        PERTENEN        PROXIMID
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.500   1st Qu.:1.500   1st Qu.:2.000   1st Qu.:1.500
##  Median :2.000   Median :2.000   Median :3.000   Median :2.000
##  Mean   :2.429   Mean   :2.286   Mean   :2.714   Mean   :1.714
##  3rd Qu.:3.500   3rd Qu.:3.000   3rd Qu.:3.500   3rd Qu.:2.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :2.000
##     FORMALID
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.143
##  3rd Qu.:2.500
##  Max.   :3.000
```

```
#Usamos la función summary para observar la frecuencia de las categorías en cada variable.
```

Para este ejemplo utilizaremos las variables correspondientes como son la intensidad, frecuencia, pertenecía, proximidad y formalidad que hacen referencia a la intensidad de la relación social. Vemos que la base de datos tiene 5 variables y para cada una 7 observaciones.

Para el análisis de MCA se utilizarán dos paqueterías especiales, una es FactoMineR()y la otra es factoextra(), y esta última se apoya en la librería ggplot(). En caso de no haberlas instalado previamente, esto se puede hacer con el siguiente comando:

```
#install.packages(c("FactoMineR", "factoextra", "tidyverse", "naniar", "corrplot","psych"))

# Y una vez que han sido instaladas, se deben activar con el comando library().
library(FactoMineR)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(factoextra)
library(naniar)
library(corrplot)
library(psych)
```

Alfa de Cronbach.

La adecuación se analiza con el Alfa de Cronbach, esta es una medida de consistencia interna de un cuestionario o escala. Evalúa cuánto están relacionadas entre sí las variables de una misma base de datos. Se calcula a partir de las correlaciones entre las variables, proporcionando un índice que varía entre 0 y 1. Valores más altos indican una mayor consistencia interna, siendo generalmente aceptable un valor de 0.7 o superior.

Para calcular el Alfa de Cronbach se utiliza la función "alpha" del paquete psych.

```
# data<-data.frame(datos)
# resultado <- alpha(data)
# print(resultado) #Mostrar el resultado
```

raw_alpha <dbl>	std.alpha <dbl>	G6(smc) <dbl>	average_r <dbl>	S/N <dbl>	ase <dbl>	mean <dbl>
0.872006	0.8708256	0.959791	0.5741591	6.741474	0.05688057	2.257143

1 row | 1-8 of 9 columns

Figure 28: "RESULTADO"

El objeto resultado contendrá la estimación del Alfa de Cronbach, así como otros indicadores útiles. Un alfa de Cronbach mayor a 0.7 generalmente se considera aceptable, pero el umbral puede variar dependiendo del contexto de tu investigación. Cuando imprimes resultado, verás una salida que incluye:

raw_alpha: El valor del Alfa de Cronbach.

std.alpha: El Alfa de Cronbach estandarizado.

En este ejemplo, el valor de Alfa de Cronbach es 0.872006, lo cual indica una buena consistencia interna de los ítems en el cuestionario. Además, puedes ver cómo cambiaría el Alfa si se eliminara alguna variable en particular.

Ahora se solicita una revisión a la estructura de las variables de interés, para ello se usa el comando str().

```
str(datos)
```

```
## 'data.frame':    7 obs. of  5 variables:
##  $ INTENSID: num  1 2 1 4 4 3 2
##   ..- attr(*, "value.labels")= Named num [1:4] 4 3 2 1
##   .. ..- attr(*, "names")= chr [1:4] "alta" "moderada" "baja" "ligera"
##  $ FRECUENC: num  1 2 1 2 4 3 3
##   ..- attr(*, "value.labels")= Named num [1:4] 4 3 2 1
##   .. ..- attr(*, "names")= chr [1:4] "frecuente" "no frecuente" "no recurrente" "ligera"
##  $ PERTENEN: num  1 2 2 4 4 3 3
##   ..- attr(*, "value.labels")= Named num [1:4] 4 3 2 1
##   .. ..- attr(*, "names")= chr [1:4] "alto" "variable" "ligero" "ninguno"
##  $ PROXIMID: num  2 2 1 2 2 1 2
##   ..- attr(*, "value.labels")= Named num [1:2] 2 1
##   .. ..- attr(*, "names")= chr [1:2] "cercano" "distante"
##  $ FORMALID: num  2 2 1 3 3 2 2
##   ..- attr(*, "value.labels")= Named num [1:3] 3 2 1
##   .. ..- attr(*, "names")= chr [1:3] "informal" "formal" "sin relaci\xf3n"
##  - attr(*, "variable.labels")= Named chr [1:5] "intensidad interacci\xf3n" "frecuencia interacci\xf3
##   ..- attr(*, "names")= chr [1:5] "INTENSID" "FRECUENC" "PERTENEN" "PROXIMID" ...
```

Dicho resumen permite identificar como están evaluadas las variables:

La variable INTENSID se evalúa del 1 al 4 como "alta", "moderada", "baja", "ligera", FRECUENC se evalúa del 1 al 4 como "frecuente" ,"no frecuente" ,"no recurrente" ,"ligera", PERTENEN se evalúa del 1 al 4 como "alto" ,"variable", "ligero", "ninguno", PROXIMID se evalúa del 1 al 2 como "cercano" o "distante" y FORMALID se evalúa del 1 al 3 como "informal" ,"formal","sin relación"

Además podemos ver que las variables fueron detectadas por R del tipo integer (enteros), pero es necesario transformarlas a otro formato tipo factor para incluirlas en el análisis MCA. Para ello se utiliza el comando factor().

```
datos $ INTENSID <- factor(datos $ INTENSID)
datos $ FRECUENC <- factor(datos $ FRECUENC)
datos $ PERTENEN <- factor(datos $ PERTENEN)
datos $ PROXIMID <- factor(datos $ PROXIMID)
datos $ FORMALID <- factor(datos $ FORMALID)
```

A continuación se inicia el análisis de MCA en R, para lo que se usará inicialmente el comando MCA(), que es parte de la librería FactoMiner().

Su estructura es: MCA(X, ncp = , graph = TRUE)

Los argumentos consisten en:

X: un data frame con n filas (individuos) y p columnas (variables categóricas).

ncp: número de dimensiones a guardar en los resultados finales.

graph: a partir de un valor lógico (TRUE, FALSE) se indica si se desea generar la gráfica correspondiente.

El análisis MCA se realizará solo sobre los individuos y variables activas o de interés, ubicadas en el data frame datos.

```
acm <- MCA(datos, graph = FALSE)
```

El resultado generado se ha guardado en un objeto denominado acm, y que consiste en una lista que contiene información diversa, correspondiente tanto a listas y matrices. Y para darle un vistazo a su contenido se usa el comando print().

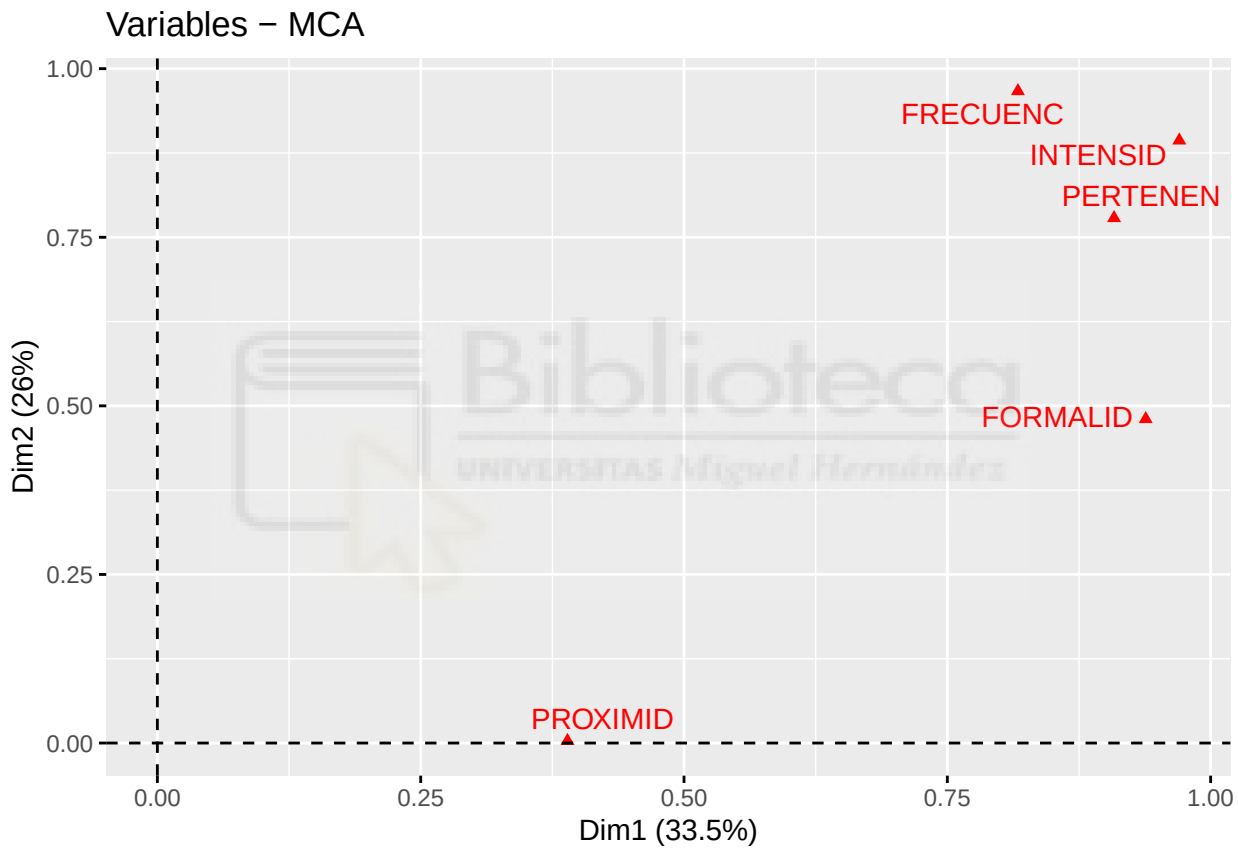
```
print(acm)
```

```
## **Results of the Multiple Correspondence Analysis (MCA)**
## The analysis was performed on 7 individuals, described by 5 variables
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$var"            "results for the variables"
## 3  "$var$coord"      "coord. of the categories"
## 4  "$var$cos2"       "cos2 for the categories"
## 5  "$var$contrib"    "contributions of the categories"
## 6  "$var$v.test"     "v-test for the categories"
## 7  "$var$eta2"       "coord. of variables"
## 8  "$ind"            "results for the individuals"
## 9  "$ind$coord"      "coord. for the individuals"
## 10 "$ind$cos2"       "cos2 for the individuals"
## 11 "$ind$contrib"    "contributions of the individuals"
## 12 "$call"           "intermediate results"
## 13 "$call$marge.col" "weights of columns"
## 14 "$call$marge.li"  "weights of rows"
```

Siguiendo la lógica del análisis que existe en el Análisis de Componentes Principales, que permite "reducir" las dimensiones de un data frame a partir de generar nuevos ejes o componentes que sirven a manera de "resumen" de las variables cuantitativas originales, en el análisis ACM (Análisis de Correspondencias Múltiple) también es posible construir dichos componentes o ejes a partir de variables categóricas.

Para visualizar la correlación entre variables y las dimensiones principales de ACM:

```
fviz_mca_var(acm, choice = "mca.cor",
             repel = TRUE,
             ggtheme = theme_grey())
```



Si observamos el siguiente gráfico vemos que las más relacionadas son las que están más cerca (intensid. y pertenen.) y las que están menos relacionadas son las más lejanas, como por ejemplo, (proximid. y frecuenc.)

Una vez que se generan los nuevos componentes, es importante identificar la capacidad explicativa del total de los casos que cada una proporciona. Para ello es importante revisar la proporción de varianzas que "retiene" cada una de estas dimensiones o ejes. Y puede ser extraído a partir de la función get_eigenvalue() de la siguiente manera:

```
eig_val <- factoextra::get_eigenvalue(acm)
head(eig_val)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 0.80465175        33.527156                    33.52716
## Dim.2 0.62452396        26.021832                    59.54899
## Dim.3 0.39719776        16.549907                    76.09889
## Dim.4 0.33385522        13.910634                    90.00953
## Dim.5 0.15548675         6.478614                    96.48814
## Dim.6 0.08428457         3.511857                   100.00000
```

En la tabla anterior se muestran del lado de las columnas los componentes o ejes nuevos, resultados del análisis ACM (Análisis de Correspondencias Múltiple), mientras que en la primer columna se muestran los eigenvalores o el tamaño de las varianzas que explica cada uno, mientras que en la segunda columna se muestra el porcentaje de la varianza total que es explicado por cada eje o dimensión. En la tercer columna se muestra el porcentaje de varianza acumulado.

En resumen, el Alfa de Cronbach y el ACM pueden utilizarse en conjunto para asegurar la fiabilidad de los datos categóricos y luego explorar visualmente las relaciones entre estos datos.

4.6. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Analizar - Reducción de dimensiones - Escalamiento óptimo - Marcamos "algunas variables no son nominales múltiples" (ya que no son nominales, son ordinales) - Pasamos todas las variables - En Gráfico seleccionamos puntos de objetos y diagrama de dispersión biespacial - En categorías pasamos todas las variables a categorías conjuntas.

Resumen del modelo

Dimensión	Alfa de Cronbach	Varianza contabilizada para Total (autovalor)	Varianza contabilizada para % de varianza
1	,882	3,400	68,008
2	,306	1,323	26,469
Total	,985[a]	4,724	94,478

a. Se utiliza el total de alfa de Cronbach en el autovalor total.

Figure 29: "RESUMEN ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLE"

La tabla de resumen del modelo, permite observar que se crearon 2 dimensiones. El autovalor muestra la proporción de información del modelo que es explicada por cada dimensión; permite analizar la importancia de cada una de ellas.

Se puede observar que la primera dimensión es más importante para el modelo que la segunda. A su vez, la primera explica más inercia (0,882) que la segunda (0,306), a mayor dependencia entre variables, mayor inercia. Esto quiere decir que las categorías presentan mayor dispersión de varianza en la dimensión 1, sin embargo ambas dimensiones tienen un valor similar de inercia.

El alfa de Cronbach indica también qué tan correlacionadas están las variables observables que componen la base de datos, por lo que ambos valores (alfa de Cronbach e inercia) tienen una relación directa. Miramos en la tabla resumen si el Alfa de Cronbach Total es más grande que 0.7, en este caso, (0.900), entonces aceptamos la relación entre las variables cualitativas.

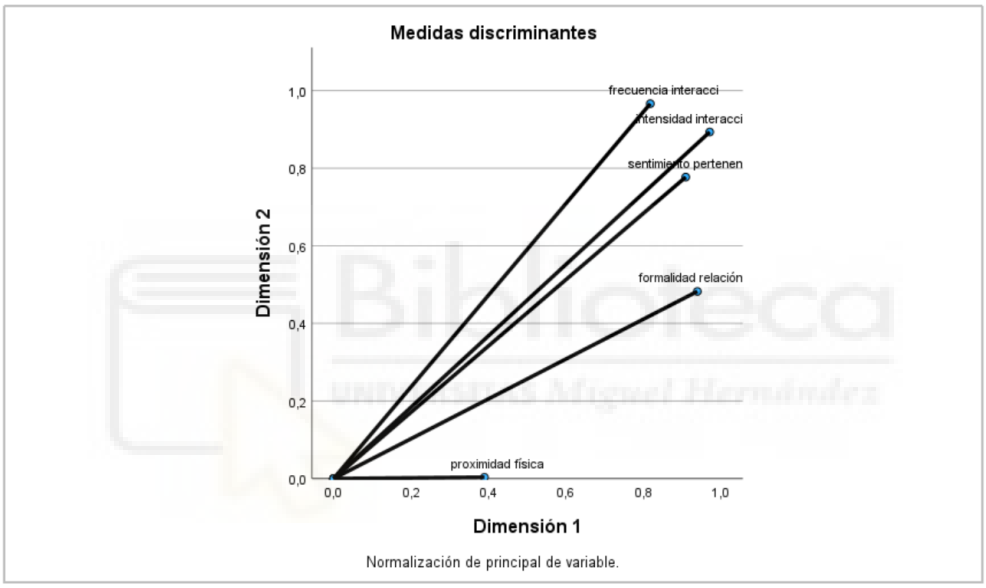


Figure 30: "GRÁFICO DEL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLE"

Si observamos el siguiente gráfico vemos que las más relacionadas son las que están más cerca (intensidad, interacción y sentimiento pertenen.) y están menos relacionadas las más lejanas (proximidad física y frecuencia intereacci)

Variables transformadas de correlaciones

	intensidad interacción	frecuencia interacción	sentimiento pertenencia	proximidad física	formalidad relación
intensidad interacción	1,000	,977	,805	,370	,773
frecuencia interacción	,977	1,000	,795	,331	,723
sentimiento pertenencia	,805	,795	1,000	-,118	,307
proximidad física	,370	,331	-,118	1,000	,685
formalidad relación	,773	,723	,307	,685	1,000
Dimensión	1	2	3	4	5
Autovalor	3,400	1,323	,200	,066	,010

Figure 31: "CORRELACIONES DEL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLE"

Aqui podemos ver "¿Qué dos variables están más correlacionadas?", vemos que ntensidad con frecuencia (0.977) y pertenencia con intensidad (0.805), sólo miramos las correlaciones más altas, si nos fijamos también en el gráfico de dispersión vemos que son las parejas que están más cerca entre ellas.

5. ESCALAMIENTO ÓPTIMO Y MULTIDIMENSIONAL.

Este capítulo no se dedica a una sola técnica sino a un conjunto de ellas. Se llama "Escalamiento" a cualquier técnica que construye una escala de medida. Se dice que el "Escalamiento" sirve para ilustrar opiniones.

El Escalamiento Óptimo es el nombre con el que se conoce al conjunto de técnicas que reducen la dimensión de una base de datos que incluye variables cualitativas, esta técnica se basa en la transformación de datos cualitativos en cuantitativos de manera que se maximice la información que se puede proporcionar en un análisis. El caso en que todas son nominales coincide con el Análisis de Correspondencias Simple o Múltiple. Mientras que el Escalamiento Multidimensional es la técnica de análisis de variables cualitativas que analiza matrices cuadradas de similitudes o disimilitudes.

En resumen, el escalamiento óptimo y multidimensional son herramientas poderosas para el análisis y visualización de datos de gran dimensión, facilitando la interpretación y exploración de estructuras complejas en los datos.

Un ejemplo de estudio que requiere Escalamiento Óptimo es el análisis de Escalas de Likert. Un ejemplo de estudio que requiere Escalamiento Multidimensional es el análisis de una matriz de parecidos.

Un requisito de este análisis sería que la base de datos debe contener variables cualitativas y que no todas sean nominales.

4.1 ESCALAMIENTO ÓPTIMO.

Como bien hemos dicho el Escalamiento Óptimo coincide con el Análisis de Correspondencias Simple o Múltiple.

Ejercicio 6. Fichero 10-4.sav. Este fichero "10-4.sav" muestra los resultados de una encuesta con afirmaciones de Likert. El sexo es una variable etiqueta.

Leemos la base de datos:

```
library(foreign)
datos <- read.spss("10-4.sav",
                   use.value.labels = F,
                   to.data.frame=TRUE)
summary(datos)
```

```
##      ITEM1           ITEM2           ITEM3           ITEM4           ITEM5
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000
##  Median :3.000   Median :3.00   Median :3.000   Median :3.000   Median :3.000
##  Mean   :2.627   Mean   :3.09   Mean   :3.194   Mean   :3.343   Mean   :2.955
##  3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      ITEM6           ITEM7           ITEM8           ITEM9           SEXO
##  Min.   :1   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:1.000
##  Median :3   Median :3.000   Median :3.00   Median :3.000   Median :2.000
##  Mean   :3   Mean   :2.896   Mean   :2.91   Mean   :2.866   Mean   :1.537
##  3rd Qu.:4   3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000
##  Max.   :5   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :2.000
```

```
#Usamos la función summary para observar la frecuencia de las categorías en cada variable.
```

Para este ejemplo utilizaremos 9 variables más la variable sexo que nos indica el enunciado que es una variable etiqueta. Esta base de datos cuenta con un total de 10 variables y 67 observaciones.

```
#install.packages(c("FactoMineR", "factoextra", "tidyverse", "naniar", "corrplot","psych"))

# Y una vez que han sido instaladas, se deben activar con el comando library().
# Utilizamos 'suppressWarnings' para cargar la librería sin mostrar warnings

suppressWarnings(suppressMessages(library(psych)))
suppressWarnings(suppressMessages(library(tidyverse)))
suppressWarnings(suppressMessages(library(FactoMineR)))
suppressWarnings(suppressMessages(library(factoextra)))
suppressWarnings(suppressMessages(library(naniar)))
suppressWarnings(suppressMessages(library(psych)))
```

Alfa de Cronbach.

La adecuación se analiza con el Alfa de Cronbach, esta es una medida de consistencia interna de un cuestionario o escala. Evalúa cuánto están relacionadas entre sí las variables de una misma base de datos. Se calcula a partir de las correlaciones entre las variables, proporcionando un índice que varía entre 0 y 1. Valores más altos indican una mayor consistencia interna, siendo generalmente aceptable un valor de 0.7 o superior.

Para calcular el Alfa de Cronbach se utiliza la función "alpha" del paquete psych.

```
# library(psych)
# Crear un data frame de la base de datos
# data <- data.frame(datos)
# Calcular el Alfa de Cronbach
# resultado <- alpha(data)
# Mostrar el resultado
# print(resultado)
```

raw_alpha <dbl>	std.alpha <dbl>	G6(smc) <dbl>	average_r <dbl>	S/N <dbl>	ase <dbl>	mean <dbl>
-0.4061336	-0.430156	-0.2016743	-0.03101027	-0.3007756	0.2524874	2.841791

1 row | 1-8 of 9 columns

Figure 32: "RESULTADO"

En este ejemplo, el valor de Alfa de Cronbach es -0.41, lo cual no indica una buena consistencia interna de los ítems en el cuestionario.

Con el comando str(), se obtendría una revisión de la estructura de las variables de interés. Dicho resumen permite identificar como están evaluadas las variables, todas se evalúan del 1 al 5, excepto la variable sexo que es evaluada como 1 o 2.

Además podemos ver que las variables fueron detectadas por R del tipo integer (enteros), pero es necesario transformarlas a otro formato tipo factor para incluirlas en el análisis ACM (Análisis de Correspondencias Múltiple). Para ello se utiliza el comando factor().

```
datos $ ITEM1 <- factor(datos $ ITEM1)
datos $ ITEM2 <- factor(datos $ ITEM2)
datos $ ITEM3 <- factor(datos $ ITEM3)
datos $ ITEM4 <- factor(datos $ ITEM4)
datos $ ITEM5 <- factor(datos $ ITEM5)
datos $ ITEM6 <- factor(datos $ ITEM6)
datos $ ITEM7 <- factor(datos $ ITEM7)
datos $ ITEM8 <- factor(datos $ ITEM8)
datos $ ITEM9 <- factor(datos $ ITEM9)
```

A continuación se inicia el análisis de MCA en R, para lo que se usará inicialmente el comando MCA(), que es parte de la librería FactoMiner().

El análisis MCA se realizará solo sobre los individuos y variables activas o de interés, ubicadas en el data frame datos.

Primero debemos eliminar la columna SEXO ya que es una variable de etiqueta, si la añadimos lo más posible es que nos diera error el comando.

```
datos2 <- subset(datos, select = -SEXO)
print(datos2)
```

```
##    ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6 ITEM7 ITEM8 ITEM9
## 1      1     5     3     4     4     2     3     2     4
## 2      3     3     4     3     3     3     2     4     5
## 3      5     2     2     2     2     2     4     3     3
## 4      4     1     5     1     4     3     5     5     2
## 5      1     5     4     5     3     4     1     1     2
## 6      2     4     4     4     2     5     3     4     3
## 7      4     3     3     5     4     3     4     2     4
## 8      3     5     3     3     3     4     2     4     3
## 9      2     2     2     2     4     2     4     2     2
## 10     1     1     2     4     2     3     3     2     3
## 11     3     2     4     3     3     4     4     3     4
## 12     5     3     5     4     2     4     3     2     2
## 13     1     2     3     3     5     3     2     3     1
## 14     1     1     5     4     4     3     4     1     4
## 15     3     4     3     3     2     2     2     2     3
## 16     4     2     5     4     3     2     5     4     2
## 17     5     1     2     3     3     2     1     5     4
## 18     2     5     3     4     3     5     5     3     3
## 19     1     3     3     3     4     3     3     2     4
```

```
## 20     4     2     4     5     4     3     2     4     2
## 21     3     4     2     3     2     2     4     1     3
## 22     2     3     3     4     2     2     3     5     4
## 23     1     2     4     2     4     3     2     4     2
## 24     4     1     2     2     4     4     3     1     4
## 25     5     1     2     4     1     3     4     5     3
## 26     3     1     3     4     5     3     2     3     2
## 27     2     4     3     3     2     3     2     2     2
## 28     1     5     4     2     2     4     5     5     5
## 29     2     1     2     4     4     4     1     1     4
## 30     3     5     2     3     3     3     3     4     3
## 31     1     2     5     2     4     3     2     2     3
## 32     3     3     4     2     1     3     4     3     2
## 33     1     4     1     4     5     2     3     2     3
## 34     3     2     1     3     1     2     2     3     1
## 35     5     4     5     2     2     3     2     2     4
## 36     3     3     3     3     3     3     1     4     3
## 37     4     2     3     5     4     3     3     1     2
## 38     1     1     4     3     3     3     4     5     3
## 39     2     4     1     4     4     1     2     2     4
## 40     1     2     4     3     1     3     4     2     1
## 41     1     5     2     3     2     5     4     5     3
## 42     2     4     2     1     3     5     3     4     2
## 43     3     5     2     4     1     3     2     1     1
## 44     1     5     4     3     2     1     5     4     3
## 45     2     3     5     4     3     3     4     2     5
## 46     4     3     5     2     4     1     4     3     3
## 47     3     4     2     3     5     3     1     5     3
## 48     4     3     5     3     3     4     2     1     3
## 49     4     3     4     3     2     4     1     5     3
## 50     1     3     2     5     4     3     2     4     3
## 51     1     2     3     4     5     3     1     2     3
## 52     4     3     2     4     3     3     4     2     4
## 53     3     5     3     4     2     3     3     1     3
## 54     4     3     2     5     3     4     2     3     3
## 55     4     5     2     4     3     2     2     4     3
## 56     4     3     5     5     2     1     1     2     4
## 57     4     5     3     2     3     3     1     5     3
## 58     3     5     4     3     2     1     5     3     2
## 59     4     2     4     2     4     3     2     1     4
## 60     2     3     5     3     2     4     3     4     2
## 61     2     3     2     4     1     2     4     5     2
## 62     1     3     2     5     4     3     2     4     1
## 63     2     3     4     2     3     1     5     1     2
## 64     3     4     2     3     1     5     4     3     2
## 65     2     3     4     5     4     3     2     1     2
## 66     2     4     3     2     3     4     3     2     1
## 67     1     3     4     5     3     5     4     3     4
```

```
acm <- MCA(datos2, graph = FALSE)
```

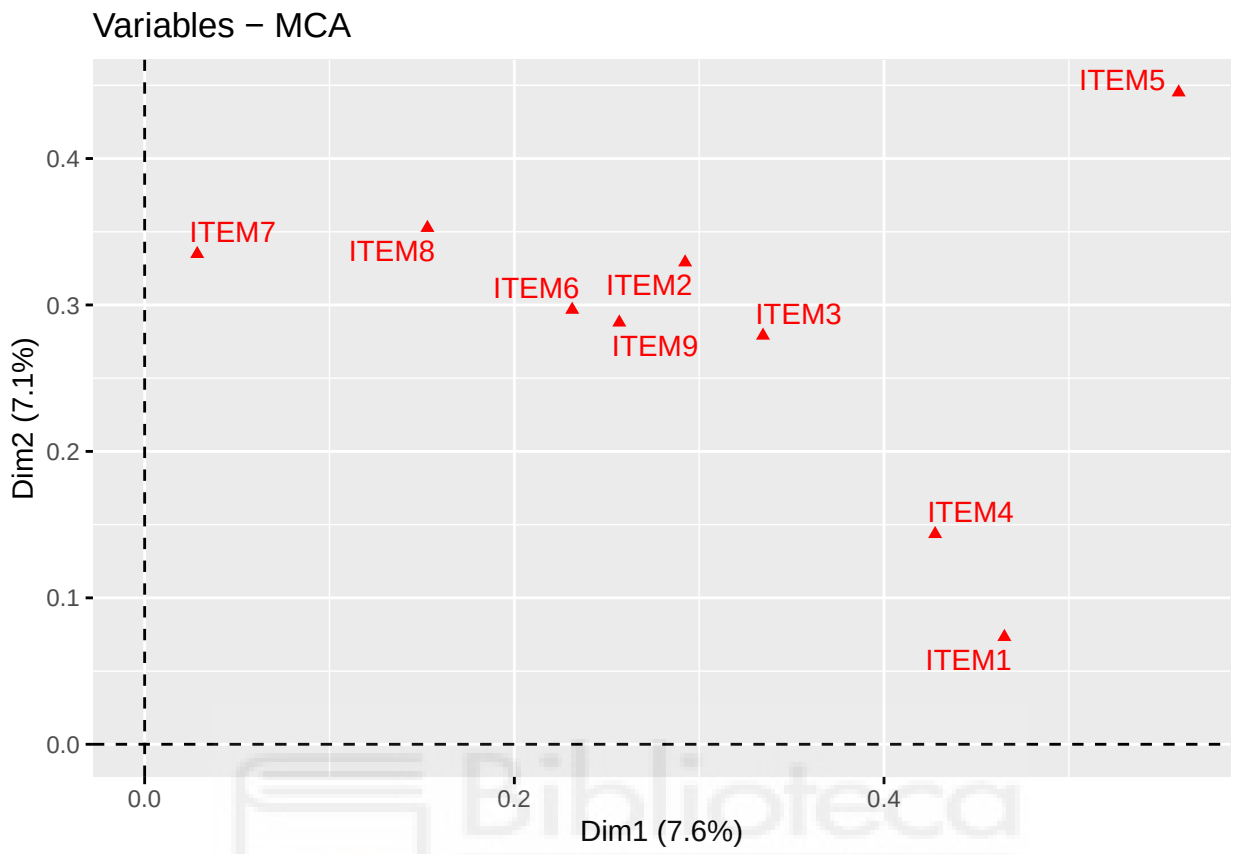
El resultado generado se ha guardado en un objeto denominado acm, y que consiste en una lista que contiene información diversa, correspondiente tanto a listas y matrices. Y para darle un vistazo a su contenido se usa el comando print().

```
print(acm)
```

```
## **Results of the Multiple Correspondence Analysis (MCA)**
## The analysis was performed on 67 individuals, described by 9 variables
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$var"            "results for the variables"
## 3  "$var$coord"      "coord. of the categories"
## 4  "$var$cos2"       "cos2 for the categories"
## 5  "$var$contrib"    "contributions of the categories"
## 6  "$var$v.test"     "v-test for the categories"
## 7  "$var$eta2"       "coord. of variables"
## 8  "$ind"            "results for the individuals"
## 9  "$ind$coord"      "coord. for the individuals"
## 10 "$ind$cos2"       "cos2 for the individuals"
## 11 "$ind$contrib"    "contributions of the individuals"
## 12 "$call"           "intermediate results"
## 13 "$call$marge.col" "weights of columns"
## 14 "$call$marge.li"  "weights of rows"
```

Para visualizar la correlación entre variables y las dimensiones principales de ACM:

```
fviz_mca_var(acm, choice = "mca.cor",
             repel = TRUE,
             ggtheme = theme_grey())
```



Si observamos el siguiente gráfico vemos que las más relacionadas son las que están más cerca (ITEM2, ITEM6 y ITEM9) y las que están menos relacionadas son las más lejanas, como por ejemplo, (ITEM1 y ITEM7).

Una vez que se generan los nuevos componentes, es importante identificar la capacidad explicativa del total de los casos que cada una proporciona. Para ello es importante revisar la proporción de varianzas que "retiene" cada una de estas dimensiones o ejes. Y puede ser extraído a partir de la función get_eigenvalue() de la siguiente manera:

```
eig_val <- factoextra::get_eigenvalue(acm);eig_val
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.30540600        7.6351500                     7.63515
## Dim.2  0.28250807        7.0627017                    14.69785
## Dim.3  0.27112123        6.7780308                    21.47588
## Dim.4  0.23559050        5.8897624                    27.36564
## Dim.5  0.22901123        5.7252806                    33.09093
## Dim.6  0.19967174        4.9917936                    38.08272
## Dim.7  0.17983822        4.4959554                    42.57867
## Dim.8  0.16213471        4.0533676                    46.63204
## Dim.9  0.15881019        3.9702548                    50.60230
## Dim.10 0.15380412        3.8451029                    54.44740
## Dim.11 0.14688674        3.6721684                    58.11957
## Dim.12 0.13944890        3.4862225                    61.60579
```

```
## Dim.13 0.13271274        3.3178184                64.92361
## Dim.14 0.12629482        3.1573704                68.08098
## Dim.15 0.11675714        2.9189285                70.99991
## Dim.16 0.11283606        2.8209015                73.82081
## Dim.17 0.10532065        2.6330163                76.45383
## Dim.18 0.10073104        2.5182761                78.97210
## Dim.19 0.09199581        2.2998952                81.27200
## Dim.20 0.08459562        2.1148904                83.38689
## Dim.21 0.07777821        1.9444553                85.33134
## Dim.22 0.07272051        1.8180128                87.14936
## Dim.23 0.06903715        1.7259287                88.87528
## Dim.24 0.06097414        1.5243535                90.39964
## Dim.25 0.05721725        1.4304313                91.83007
## Dim.26 0.05433853        1.3584634                93.18853
## Dim.27 0.04507170        1.1267925                94.31533
## Dim.28 0.03809333        0.9523331                95.26766
## Dim.29 0.03712630        0.9281576                96.19582
## Dim.30 0.03248024        0.8120059                97.00782
## Dim.31 0.03050406        0.7626014                97.77042
## Dim.32 0.02496155        0.6240387                98.39446
## Dim.33 0.01940898        0.4852245                98.87969
## Dim.34 0.01756925        0.4392312                99.31892
## Dim.35 0.01410719        0.3526797                99.67160
## Dim.36 0.01313611        0.3284027               100.00000
```

4.1.1. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Analizar - Reducción de dimensiones - Escalamiento óptimo - Marcamos "algunas variables no son nominales múltiples" (ya que no son nominales, son ordinales) - Pasamos todas las variables - En Gráfico seleccionamos puntos de objetos y diagrama de dispersión biespacial - En categorías pasamos todas las variables a categorías conjuntas.

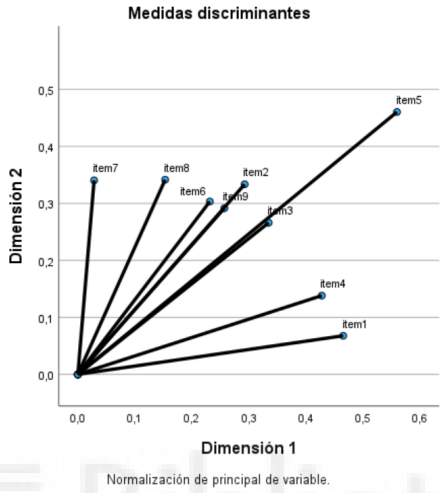


Figure 33: "GRAFICO"

4.2 ESCALAMIENTO MULTIDIMENSIONAL.

El Escalamiento Multidimensional es la técnica de análisis de variables cualitativas que analiza matrices cuadradas de similitudes o disimilitudes. Las matrices de similitudes son las que hacen corresponder valores grandes a grandes parecidos y las matrices de disimilitudes las que hacen corresponder valores grandes a pequeños parecidos. Las matrices de distancias son matrices de disimilitudes, mientras que las matrices de contingencia son las que recogen el número de coincidencias son matrices de similitudes.

El objetivo del Escalamiento Multidimensional es transformar los juicios de similitud o preferencias de un conjunto de individuos en distancias susceptibles de ser representadas en un espacio multidimensional. El requisito que existe es que la base de datos contenga variables cualitativas pero no todas nominales.

Ejercicio 7. Fichero 10-1.sav. Este fichero muestra la matriz de distancias entre las ciudades europeas (disimilitud):

Leemos los datos:

```
#install.packages('haven')
library('haven')
ciudades<-as.data.frame(read_sav("10-1.sav"))
```

	CIUDADES	ATENAS	BERLÍN	ESTOCOLM	LONDRES	MADRID	MOSCÚ	PARÍS	ROMA	VARSOVIA	VIENA
1	Atenas	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	Berlín	1774	0	NA	NA	NA	NA	NA	NA	NA	NA
3	Estoco	2371	806	0	NA	NA	NA	NA	NA	NA	NA
4	Londre	2355	919	1387	0	NA	NA	NA	NA	NA	NA
5	Madrid	2387	1855	2548	1258	0	NA	NA	NA	NA	NA
6	Moscú	2177	1565	1210	2419	3371	0	NA	NA	NA	NA
7	París	2065	871	1516	339	1048	2419	0	NA	NA	NA
8	Roma	1048	1177	1952	1419	1371	2323	1097	0	NA	NA
9	Varsov	1581	484	790	1403	2258	1129	1323	1290	0	NA
10	Viena	1274	516	1226	1210	1806	1613	1016	758	548	0

```
matriz <- matrix(c(0,1774,2371,2355,2387,2177,2065,1048,1581,1274,
                   1774,0,806,919,1855,1565,871,1177,484,516,
                   2371,806,0,1387,2548,1210,1516,1952,790,1226,
                   2355,919,1387,0,1258,2419,339,1419,1403,1210,
                   2387,1855,2548,1258,0,3371,1048,1371,2258,1806,
                   2177,1565,1210,2419,3371,0,2419,2323,1129,1613,
                   2065,871,1516,339,1048,2419,0,1097,1323,1016,
                   1048,1177,1952,1419,1371,2323,1097,0,1290,758,
                   1581,484,790,1403,2258,1129,1323,1290,0,548,
                   1274,516,1226,1210,1806,1613,1016,758,548,0
                   ), nrow = 10, byrow = TRUE)

matriz # Mostrar la matriz resultante
```

Creamos la matriz simétrica con diagonal 0:

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    0 1774 2371 2355 2387 2177 2065 1048 1581  1274
##  [2,] 1774    0  806  919 1855 1565  871 1177  484   516
##  [3,] 2371  806    0 1387 2548 1210 1516 1952  790  1226
##  [4,] 2355  919 1387    0 1258 2419  339 1419 1403  1210
##  [5,] 2387 1855 2548 1258    0 3371 1048 1371 2258  1806
##  [6,] 2177 1565 1210 2419 3371    0 2419 2323 1129  1613
##  [7,] 2065  871 1516  339 1048 2419    0 1097 1323  1016
##  [8,] 1048 1177 1952 1419 1371 2323 1097    0 1290   758
##  [9,] 1581  484  790 1403 2258 1129 1323 1290    0   548
## [10,] 1274  516 1226 1210 1806 1613 1016  758  548     0
```

Para realizar el análisis de escalamiento multidimensional utilizamos la función "cmdscale" introduciendo la matriz de disimilitudes.

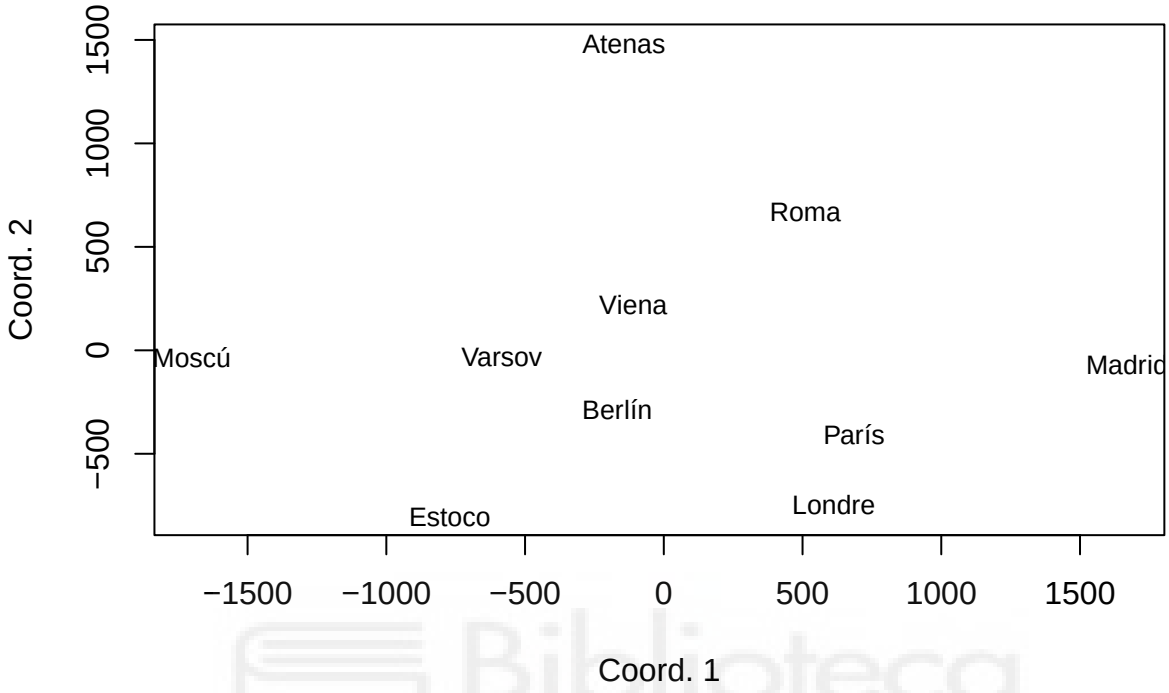
```
# Ejecutar el escalado multidimensional
modelo <- cmdscale(matriz)
modelo
```

```
##             [,1]       [,2]
##  [1,]  -143.6392 1483.49192
##  [2,]  -167.6854 -285.28100
##  [3,]  -771.3626 -801.87185
##  [4,]   612.3738 -746.05882
##  [5,]  1669.2086  -66.99807
##  [6,] -1700.9287  -35.09954
##  [7,]   685.9740 -406.99828
##  [8,]   510.1366  669.31413
##  [9,]  -584.1371  -30.15067
## [10,]  -109.9398  219.65219
```

```
# Calcular las distancias obtenidas
distancia <- dist(modelo)
distancia
```

```
##           1        2        3        4        5        6        7
## 2  1768.9364
## 3  2370.0051  794.5390
## 4  2354.2413  905.9849 1384.8616
## 5  2385.4636 1849.8181 2548.8090 1256.1940
## 6  2175.1484 1553.5205 1205.0033 2420.0892 3370.2883
## 7  2064.5124  862.2931 1509.8858  346.9568 1040.3607 2415.7014
## 8  1044.1783 1170.7666 1951.0583 1419.0606 1373.1728 2320.5621 1090.5811
## 9  1576.4366  488.3887  794.1077 1394.3324 2253.6469 1116.8026 1324.8382
## 10 1264.2889  508.2244 1216.9600 1205.9580 1802.0925 1611.2555 1013.0002
##           8        9
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9  1298.7247
## 10  765.9573  535.9707
```

```
# Representar los objetos como puntos de un mapa de dimensión dos:
rownames(ciudades) <- colnames(ciudades) <- c("Atenas", "Berlín", "Estoco", "Londre", "Madrid","Moscú",

plot(modelo, type = "n", xlab = "Coord. 1", ylab = "Coord. 2")
text(modelo[, 1], modelo[, 2], labels = rownames(ciudades), cex = 0.8)
```




```r
# Esta función nos da mucha información sobre los valores del modelo
valores <- cmdscale(distancia, eig = T)
```

Conclusiones del gráfico:

Madrid está lo más lejos de Moscú, París y Londres están situados más cerca. Si una variable está cerca del origen, no significa que esté mal explicado, solo estamos escalando las variables. No hay que concluir sobre si están bien o mal explicadas, es decir, no estamos reduciendo dimensiones.

También tendremos que decidir si la matriz es de disimilitudes o similitudes dependiendo de si la matriz significa que están cerca o lejos. En nuestro caso es de disimilitudes ya que son las distancias entre ciudades, y las distancias siempre son DISIMILITUD.

4.2.1. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Analizar - Escala - Escalamiento multidimensional (ALSCAL) - Pasamos todas las variables menos ciudades - Modelo (marcamos intervalo) - Opciones (gráfico de grupo) - Aceptar

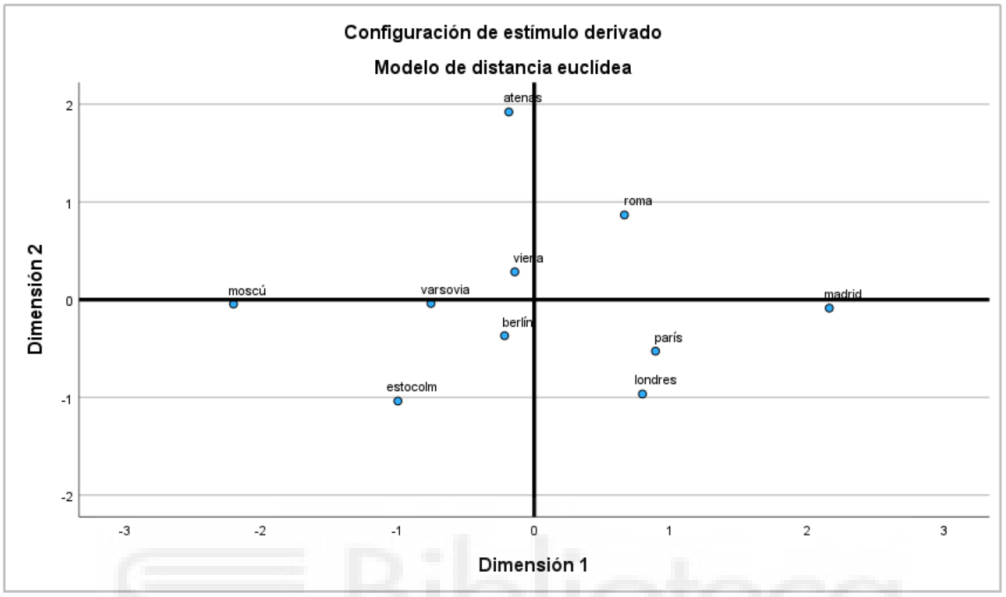


Figure 34: "MATRIZ DE DISIMILITUDES"

6. ANÁLISIS CLÚSTER.

El análisis clúster, también conocido como análisis de conglomerados, es una técnica de análisis multivariante que sirve para clasificar a los individuos en grupos homogéneos en función de las variables en la base de datos. La elección de los grupos depende del criterio. Por ejemplo, una baraja de cartas queda igual de bien clasificada si la consideramos en grupos de figuras (ases, reyes, etc.) o en palos (espadas, bastos, etc.)

Para que el análisis clúster dé resultados interesantes es necesario que las variables que usemos no estén muy correladas entre sí, es decir, si están muy correladas obtendría información muy similar con las distintas variables pero si están poco correladas tendré información más variada.

El análisis clúster es la continuación natural de lo que hemos estudiado hasta ahora, porque sigue a un análisis factorial o a uno de correspondencias o a cualquiera de reducción de la dimensión. Clasificación de los métodos clúster:

Según el tipo de variable y el tipo de base de datos el análisis clúster se realiza según un algoritmo u otro. La clasificación más general de estos métodos distingue entre:

-Métodos jerárquicos

-Métodos no jerárquicos

Los métodos jerárquicos sirven para variables de cualquier tipo y no necesitan conocer el número de clúster a priori. Se usan cuando el número de individuos es inferior a 120. Se llaman así porque el modo de construir los clúster establece jerarquías entre los que están a la misma distancia y los que están más lejos.

En este método se construye un árbol de clusters, llamado dendrograma, donde cada nivel del árbol representa una partición diferente de los datos. Acerca del dendograma:

-Los empates se deshacen de modo arbitrario

-Cada línea horizontal en el dendograma define una partición en clúster.

-Para decidir el número de clúster o bien se asume una distancia inadmisible o bien se busca un codo.

-La homogeneidad de los datos será mayor cuanto más baja sea la línea horizontal más alta del dendograma.

Los métodos no jerárquicos sólo sirven para variables cuantitativas y requieren conocer el número de clúster. Pueden aborda problemas con más de 120 individuos. Mientras que en este otro método se particiona directamente los datos en un número predefinido de clusters.

Un método no jerárquico muy popular es el método de las k-medias que consiste en asignar cada objeto al cluster cuyo centroide (promedio) es más cercano.

En resumen, el análisis cluster es una técnica estadística utilizada para agrupar un conjunto de objetos en grupos o clusters de tal manera que los objetos dentro de cada grupo sean más similares entre sí que con los objetos de otros grupos. Esta similitud se basa en una serie de características o variables que describen a los objetos. Sus objetivos principales son identificar grupos homogéneos dentro de una población heterogénea y maximizar la similitud dentro de los grupos y minimizar la similitud entre los grupos.

5.1. Ejercicio 8. Realiza un Análisis Clúster no jerárquico de los países en el fichero MUNDO a partir de las variables población, urbana y densidad.

Instalación de los paquetes necesarios para el Análisis Clúster:

```
#install.packages("tidyverse")
library(tidyverse)  # data manipulation
#install.packages("cluster")
library(cluster) # clustering algorithms
#install.packages("factoextra")
library(factoextra) # clustering algorithms & visualization
#install.packages('haven')
library('haven') # Leer datos en formato SPSS
```

Leemos los datos

```
mundo <- as.data.frame(read_sav("mundo.sav"))
summary(mundo)
```

```
##      país               poblac          densidad          urbana
##  Length:109         Min.   :    256   Min.   :   2.3   Min.   :  5.00
##  Class :character   1st Qu.:   5100   1st Qu.:  29.0   1st Qu.: 40.75
##  Mode  :character   Median :  10400   Median :  64.0   Median : 60.00
##                     Mean   :  47724   Mean   : 203.4   Mean   : 56.53
##                     3rd Qu.:  35600   3rd Qu.: 126.0   3rd Qu.: 75.00
##                     Max.   :1205200   Max.   :5494.0   Max.   :100.00
##                                                        NA's   :1
##     relig              espvidaf        espvidam        alfabet
##  Length:109         Min.   :43.00   Min.   :41.00   Min.   : 18.00
##  Class :character   1st Qu.:67.00   1st Qu.:61.00   1st Qu.: 63.00
##  Mode  :character   Median :74.00   Median :67.00   Median : 88.00
##                     Mean   :70.16   Mean   :64.92   Mean   : 78.34
##                     3rd Qu.:78.00   3rd Qu.:72.00   3rd Qu.: 98.00
##                     Max.   :82.00   Max.   :76.00   Max.   :100.00
##                                                     NA's   :2
##     inc_pob          mortinf          pib_cap          región
##  Min.   :-0.300   Min.   :  4.00   Min.   :  122   Min.   :1.00
##  1st Qu.: 0.520   1st Qu.:  9.30   1st Qu.: 1000   1st Qu.:2.00
##  Median : 1.800   Median : 27.70   Median : 2995   Median :4.00
##  Mean   : 1.682   Mean   : 42.31   Mean   : 5860   Mean   :3.55
##  3rd Qu.: 2.680   3rd Qu.: 63.00   3rd Qu.: 7467   3rd Qu.:5.00
##  Max.   : 5.240   Max.   :168.00   Max.   :23474   Max.   :6.00
##
##     calorías         sida            tasa_nat       tasa_mor
##  Min.   :1667   Min.   :     0.0   Min.   :10.00   Min.   : 2.000
##  1st Qu.:2256   1st Qu.:    48.2   1st Qu.:14.00   1st Qu.: 6.848
##  Median :2653   Median :   386.5   Median :25.00   Median : 9.000
##  Mean   :2754   Mean   :  7914.3   Mean   :25.92   Mean   : 9.557
##  3rd Qu.:3226   3rd Qu.:  3175.5   3rd Qu.:35.00   3rd Qu.:11.000
##  Max.   :3825   Max.   :411907.0   Max.   :53.00   Max.   :24.000
```

```
##  NA's   :34     NA's   :3                          NA's   :1
##     tasasida           log_pib         logtsida         nac_def
##  Min.   :  0.0000   Min.   :2.086   Min.   :0.0000   Min.   : 0.9231
##  1st Qu.:  0.2763   1st Qu.:3.000   1st Qu.:0.7731   1st Qu.: 1.5417
##  Median :  5.0057   Median :3.476   Median :1.3799   Median : 2.6667
##  Mean   : 24.3794   Mean   :3.422   Mean   :1.3800   Mean   : 3.2035
##  3rd Qu.: 22.5859   3rd Qu.:3.873   3rd Qu.:1.8647   3rd Qu.: 4.1750
##  Max.   :326.7473   Max.   :4.371   Max.   :3.1830   Max.   :14.0000
##  NA's   :3                          NA's   :3        NA's   :1
##     fertilid        log_pob         cregrano        alfabmas
##  Min.   :1.300   Min.   :2.408   Min.   : 0.00   Min.   : 28.00
##  1st Qu.:1.880   1st Qu.:3.708   1st Qu.: 6.00   1st Qu.: 63.00
##  Median :3.050   Median :4.017   Median :13.50   Median : 87.00
##  Mean   :3.563   Mean   :4.114   Mean   :17.98   Mean   : 78.73
##  3rd Qu.:5.000   3rd Qu.:4.551   3rd Qu.:26.75   3rd Qu.: 96.00
##  Max.   :8.190   Max.   :6.081   Max.   :77.00   Max.   :100.00
##  NA's   :2                       NA's   :3       NA's   :24
##     alfabfem          clima         Zurbana          Zespvidaf
##  Min.   :  9.00   Min.   :1.00   Min.   :-2.1290   Min.   :-2.5687
##  1st Qu.: 45.00   1st Qu.:5.00   1st Qu.:-0.6519   1st Qu.:-0.2985
##  Median : 71.00   Median :5.00   Median : 0.1435   Median : 0.3636
##  Mean   : 67.26   Mean   :5.71   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 93.00   3rd Qu.:8.00   3rd Qu.: 0.7632   3rd Qu.: 0.7420
##  Max.   :100.00   Max.   :9.00   Max.   : 1.7961   Max.   : 1.1203
##  NA's   :24       NA's   :2      NA's   :1
##    Zespvidam          Zalfabet          Zinc_pob          Zmortinf
##  Min.   :-2.5793   Min.   :-2.6367   Min.   :-1.65535   Min.   :-1.0061
##  1st Qu.:-0.4225   1st Qu.:-0.6702   1st Qu.:-0.97063   1st Qu.:-0.8670
##  Median : 0.2246   Median : 0.4223   Median : 0.09821   Median :-0.3838
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.7638   3rd Qu.: 0.8593   3rd Qu.: 0.83304   3rd Qu.: 0.5433
##  Max.   : 1.1952   Max.   : 0.9467   Max.   : 2.97072   Max.   : 3.3007
##                    NA's   :2
##      final                FAC1_1
##  Min.   :-2.4302146   Min.   :-2.4221
##  1st Qu.:-0.5397000   1st Qu.:-0.5396
##  Median : 0.3605648   Median : 0.3600
##  Mean   :-0.0008855   Mean   : 0.0000
##  3rd Qu.: 0.7937780   3rd Qu.: 0.7937
##  Max.   : 1.1613814   Max.   : 1.1624
##  NA's   :2            NA's   :2
```

En el análisis descriptivo básico observamos que esta base de datos tiene 34 variables. Las variables país y religión son "character" lo que significa que son variables cualitativas mientras que el resto son cuantitativas y todas se encuentran en el mismo orden de magnitud.

Seleccionamos las variables que necesitamos para realizar el análisis clúster de la base de datos MUNDO:

```
datos <- subset(mundo, select = c("poblac", "urbana", "densidad"))
summary(datos)
```

```
##      poblac            urbana          densidad
##  Min.   :    256   Min.   :  5.00   Min.   :   2.3
##  1st Qu.:   5100   1st Qu.: 40.75   1st Qu.:  29.0
##  Median :  10400   Median : 60.00   Median :  64.0
##  Mean   :  47724   Mean   : 56.53   Mean   : 203.4
##  3rd Qu.:  35600   3rd Qu.: 75.00   3rd Qu.: 126.0
##  Max.   :1205200   Max.   :100.00   Max.   :5494.0
##                    NA's   :1
```

Observamos si existen valores perdidos

```
any(is.na(datos))
```

```
## [1] TRUE
```

"TRUE" lo que nos indica que sí que hay valores perdidos.

Eliminamos las filas con valores faltantes

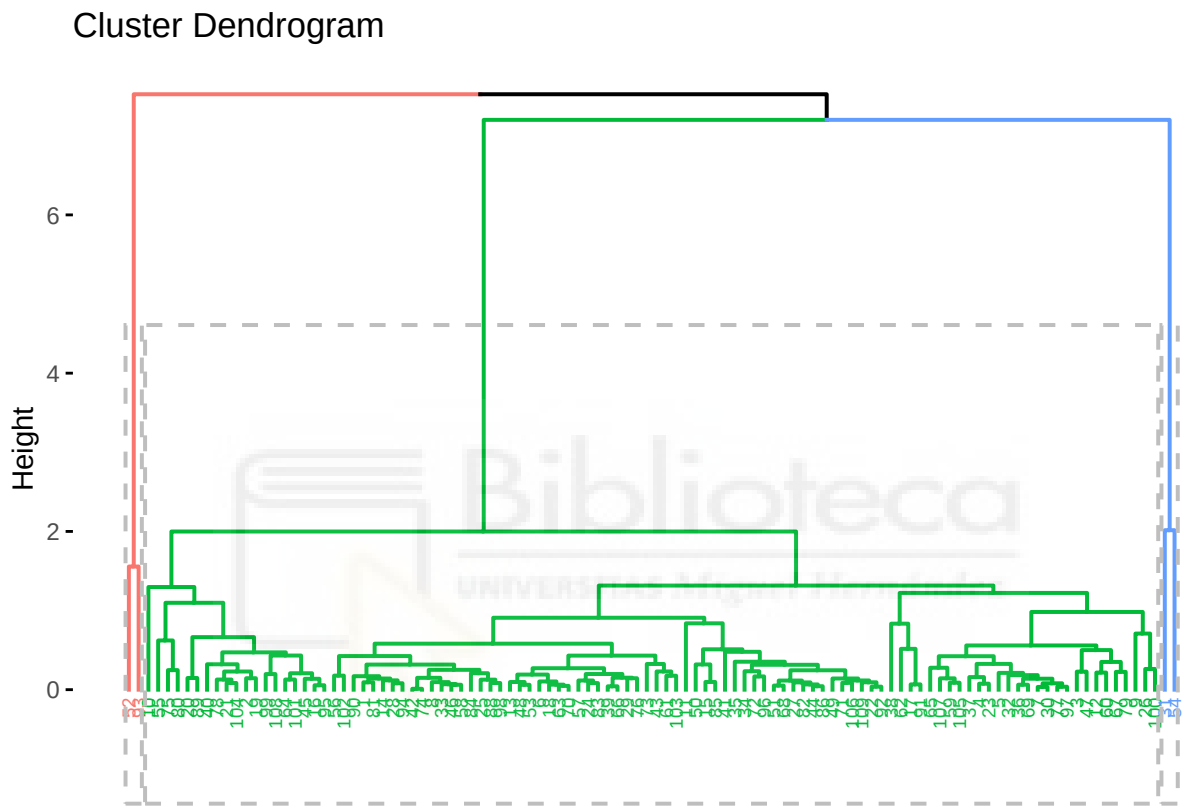
```
datos <- na.omit(datos) #eliminamos
```

Tipificamos los datos

```
G=cov.wt(datos, method='ML')$center
V=cov.wt(datos, method='ML')$cov
escala=diag(cov.wt(datos, method='ML')$cov)
Z=scale(datos,center=G,scale=sqrt(escala))
Z=as.matrix(Z)
```

Vamos a realizar una representación del dendograma:

```
#Cluster jerarquico (dendrogram)
hc=hclust(dist(Z,method = "euclidean"),method="average")

#Representamos el dendograma
fviz_dend(hc, cex = 0.5, k = 3, rect = TRUE)
```



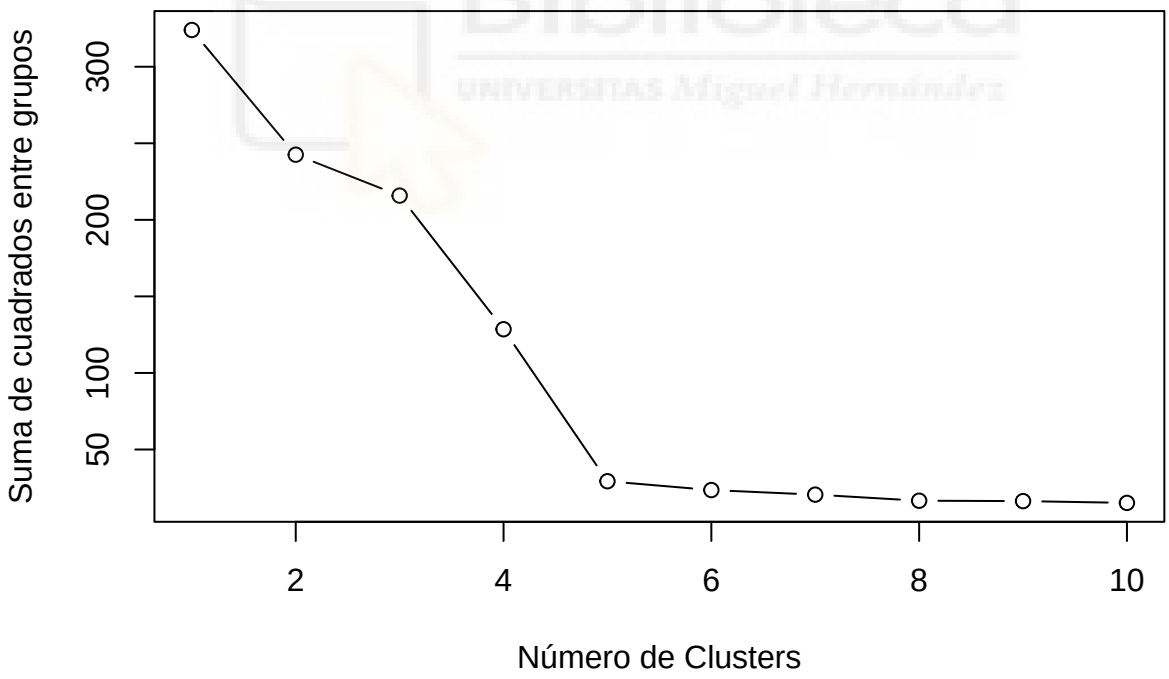
Según esta representación vemos que debebriamos dividir a los individuos en tres grupos. Para corroborar lo que hemos apreciado en el dendograma para a utilizar el Método de Elbow, este método también es conocido como 'codo' es una técnica utilizada en el análisis de clusters para determinar el número óptimo de clusters (K) en un conjunto de datos.

Existen diversos criterios para determinar el número óptimo de conglomerados a generar. A modo de ejemplo, nosotros vamos a utilizar el método de "codo" (elbow method) calcula la suma de las distancias al cuadrado de cada punto hacia su centro asignado. Este método sugiere como óptimo aquel valor de k (cantidad de conglomerados) a partir del cual añadir un conglomerado adicional apenas consigue mejoría, como sucede al pasar de K=5 a K=6.

El método de Elbow es una técnica simple y visualmente intuitiva para seleccionar el número óptimo de clusters, aunque debe usarse junto con otras técnicas y un buen entendimiento del dominio de los datos para obtener los mejores resultados.

Método de Elbow:

```
wcss <- vector()
for(i in 1:10){
  wcss[i] <- sum(kmeans(Z, i)$withinss)
}

plot(1:10, wcss, type="b", xlab="Número de Clusters", ylab="Suma de cuadrados entre grupos")
```



No siempre hay un codo claro, en algunos casos, la curva puede no tener un codo bien definido, lo que puede hacer que la determinación del número óptimo de clusters sea subjetiva. En este caso, podemos observar que existe un codo en 4 clústers.

Como la decisión que hemos tomado es subjetiva vamos a comparar el método de las K-medias con K=2, k=3 y k=4 para ver finalmente cuál es la mejor opción.

Método de las K-medias. Este algoritmo de clasificación no supervisada agrupa objetos en k grupos basándose en la mínima suma de distancias entre cada objeto y el centroide de su grupo o cluster.

Método de las K-Medias (K=2)

```
set.seed(20)
k.means <-kmeans(Z, 2, nstart = 10)
k.means
```

```
## K-means clustering with 2 clusters of sizes 106, 2
##
## Cluster means:
##         poblac      urbana   densidad
## 1  0.005623701 -0.03169753 -0.1332269
## 2 -0.298056161  1.67996897  7.0610277
##
## Clustering vector:
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1
## 102 103 104 105 106 107 108 109
##   1   1   1   1   1   1   1   1
##
## Within cluster sum of squares by cluster:
## [1] 215.258970   1.211246
##  (between_SS / total_SS =  33.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
#prototipos
k.means$centers
```

```
##         poblac      urbana   densidad
## 1  0.005623701 -0.03169753 -0.1332269
## 2 -0.298056161  1.67996897  7.0610277
```

```
k.means$cluster #clusters, es decir, grupos que se han hecho
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1
```

```
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1
## 102 103 104 105 106 107 108 109
##   1   1   1   1   1   1   1   1
```

```
k.means$totss #la suma total de cuadrados
```

```
## [1] 324
```

```
k.means$betweenss #la suma de cuadrados entre grupos
```

```
## [1] 107.5298
```

```
k.means$withinss #la suma de cuadrados intra-conglomerado
```

```
## [1] 215.258970   1.211246
```

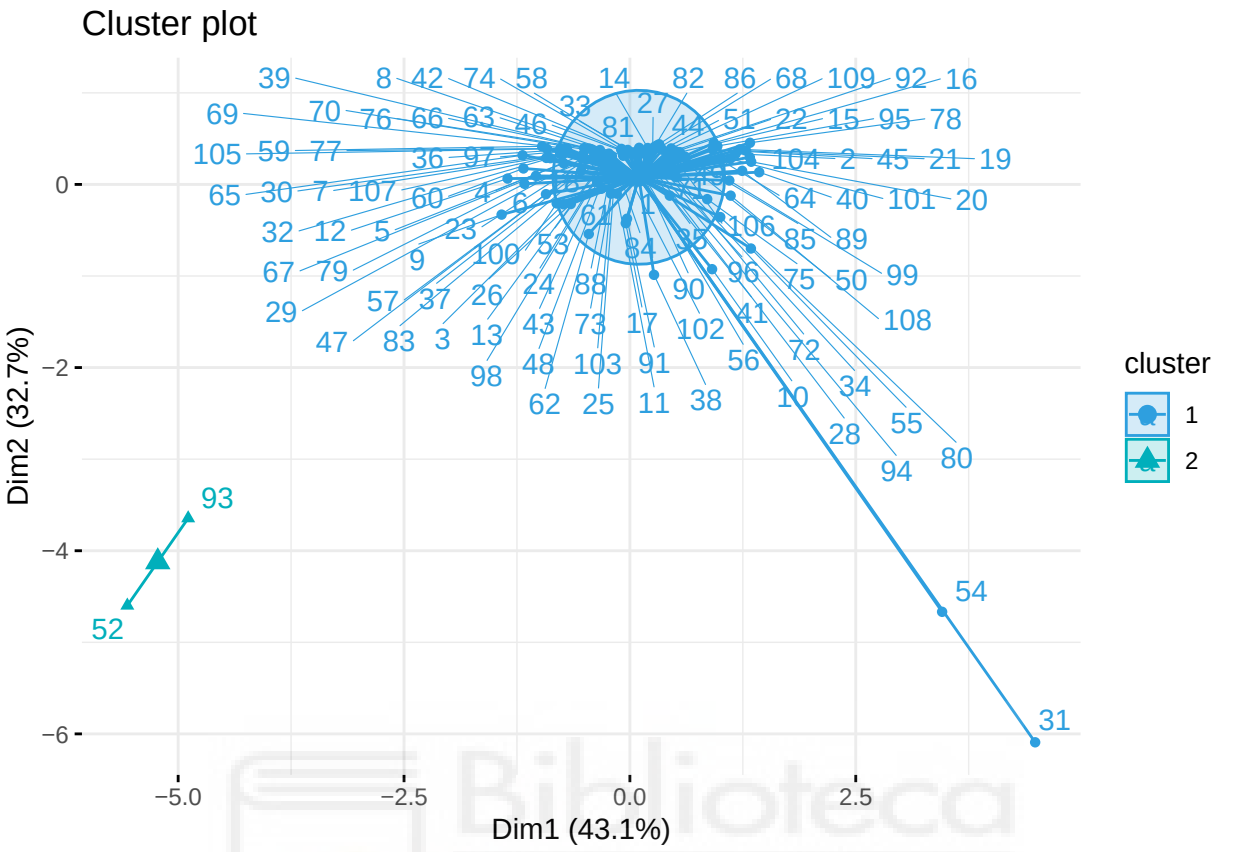
```
k.means$tot.withinss #información adicional, suma total de cuadrados dentro de cada conglomerado
```

```
## [1] 216.4702
```


Visualización de los clusters de k-means = 2

```
fviz_cluster(k.means, data = Z,
palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
ellipse.type = "euclid", # Concentration ellipse
star.plot = TRUE, # Add segments from centroids to items
repel = TRUE, # Avoid label overplotting (slow)
ggtheme = theme_minimal()
)
```

```
## Too few points to calculate an ellipse
```



En la gráfica se representa cada clúster con un color diferente, utilizando K=2, se forman dos clúster, obtenemos que todos los individuos pertenecen al clúster 1, excepto los individuos 52 y 93 que pertenecerían al clúster 2.

Método de las K-Medias (K=3)

```
set.seed(20)
k.means <-kmeans(Z, 3, nstart = 10)
k.means
```

```
## K-means clustering with 3 clusters of sizes 2, 63, 43
##
## Cluster means:
##       poblac     urbana   densidad
## 1 -0.2980562  1.6799690  7.0610277
## 2 -0.1292413  0.6554169 -0.1356909
## 3  0.2032167 -1.0384000 -0.1296169
##
## Clustering vector:
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   2   3   2   2   2   2   2   2   2   3   3   2   2   2   3   3   2   2   3   3
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   3   3   2   2   2   2   3   2   2   2   3   2   2   3   3   2   2   2   2   3
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   3   2   2   3   3   2   2   2   3   3   3   1   2   3   3   2   2   2   2   2
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   2   2   2   3   2   2   2   3   2   2   3   3   2   2   3   2   2   3   2   3
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   2   3   2   2   3   3   2   3   2   2   3   1   3   3   3   2   2   3   2   3
## 102 103 104 105 106 107 108 109
##   2   2   3   2   3   2   3   3
##
## Within cluster sum of squares by cluster:
## [1]   1.211246  24.642086 114.468919
##  (between_SS / total_SS =  56.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
#prototipos
k.means$centers
```

```
##       poblac     urbana   densidad
## 1 -0.2980562  1.6799690  7.0610277
## 2 -0.1292413  0.6554169 -0.1356909
## 3  0.2032167 -1.0384000 -0.1296169
```

```
#Clusters #grupos que se han hecho
k.means$cluster
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   2   3   2   2   2   2   2   2   2   3   3   2   2   2   3   3   2   2   3   3
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
```

```
##   3   3   2   2   2   2   3   2   2   2   3   2   2   3   3   2   2   2   2   3
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   3   2   2   3   3   2   2   2   3   3   3   1   2   3   3   2   2   2   2   2
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   2   2   2   3   2   2   2   3   2   2   3   3   2   2   3   2   2   3   2   3
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   2   3   2   2   3   3   2   3   2   2   3   1   3   3   3   2   2   3   2   3
## 102 103 104 105 106 107 108 109
##   2   2   3   2   3   2   3   3
```

```
k.means$totss
```

```
## [1] 324
```

```
k.means$betweenss
```

```
## [1] 183.6777
```

```
k.means$withinss
```

```
## [1]   1.211246  24.642086 114.468919
```

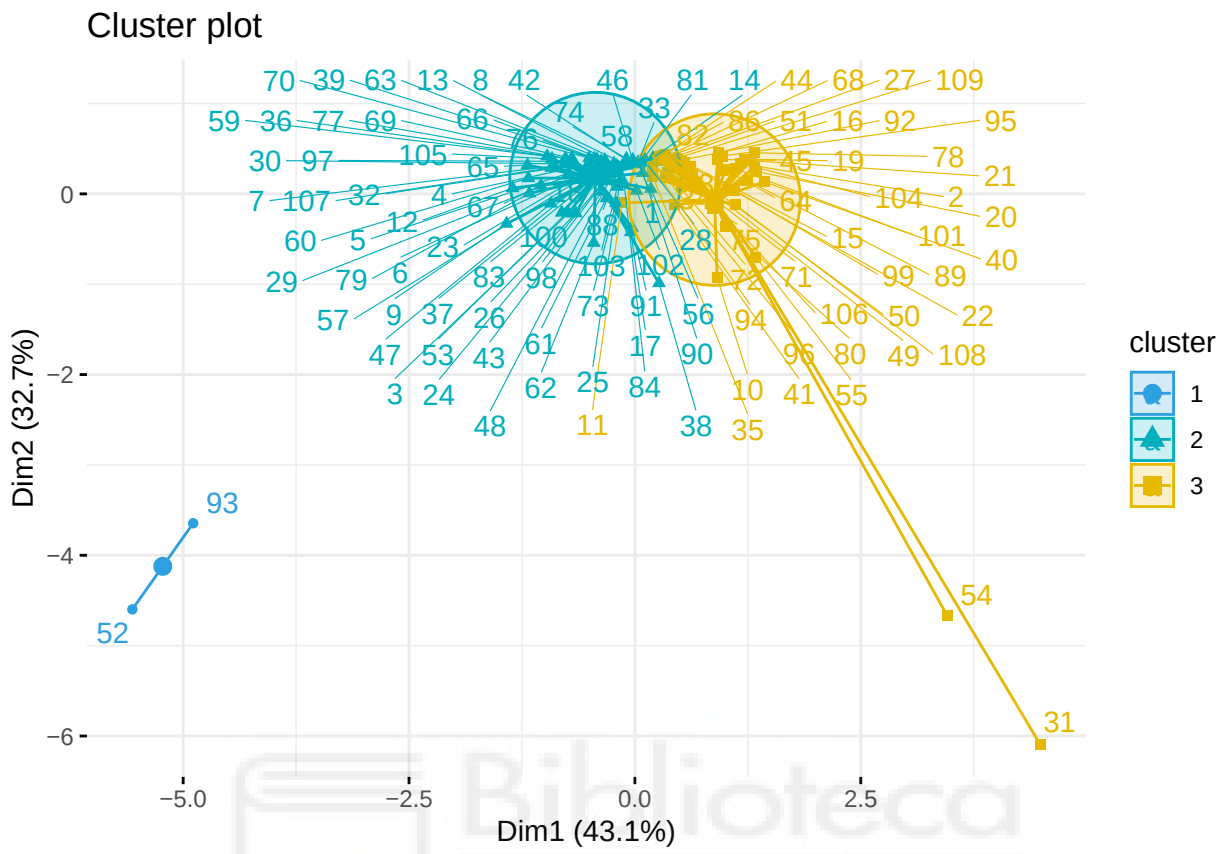
```
k.means$tot.withinss
```

```
## [1] 140.3223
```

Visualización de los clusters de k-means = 3

```
fviz_cluster(k.means, data = Z,
palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
ellipse.type = "euclid",# Concentration ellipse
star.plot = TRUE, # Add segments from centroids to items
repel = TRUE, # Avoid label overplotting (slow)
ggtheme = theme_minimal()
)
```

```
## Too few points to calculate an ellipse
```



Utilizando K=3, se forman 3 clúster, observamos que el clúster principal de antes se divide en clúster 2 y 3, mientras que el clúster 1 sigue siendo igual. Los países que pertenecen al clúster 1 son Hong Kong y Singapur.

Método de las K-Medias (K=4)

```
set.seed(20)
k.means <-kmeans(Z, 4, nstart = 10)
k.means
```

```
## K-means clustering with 4 clusters of sizes 2, 2, 60, 44
##
## Cluster means:
##       poblac     urbana      densidad
## 1  6.8878974 -1.2671832 -0.0008523776
## 2 -0.2980562  1.6799690  7.0610277230
## 3 -0.1210676  0.6982767 -0.1304792808
## 4 -0.1344460 -0.9709585 -0.1429907691
##
## Clustering vector:
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   3   4   3   3   3   3   3   3   3   4   4   3   3   4   4   4   3   3   4   4
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   4   4   3   3   3   3   4   4   3   3   1   3   3   4   4   3   3   3   3   4
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   4   3   3   4   4   3   3   3   4   4   4   2   3   1   4   3   3   3   3   3
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   3   3   3   4   3   3   3   4   3   3   4   4   3   3   4   3   3   4   3   4
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   4   4   3   3   4   4   3   4   3   3   4   2   4   4   4   3   3   4   3   4
## 102 103 104 105 106 107 108 109
##   3   3   4   3   4   3   4   4
##
## Within cluster sum of squares by cluster:
## [1]  2.030908  1.211246 22.196961 20.591835
##  (between_SS / total_SS =  85.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
#prototipos
k.means$centers
```

```
##       poblac     urbana      densidad
## 1  6.8878974 -1.2671832 -0.0008523776
## 2 -0.2980562  1.6799690  7.0610277230
## 3 -0.1210676  0.6982767 -0.1304792808
## 4 -0.1344460 -0.9709585 -0.1429907691
```

```
#Clusters #grupos que se han hecho
k.means$cluster
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
```

```
##   3   4   3   3   3   3   3   3   3   4   4   3   3   4   4   4   3   3   4   4
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   4   4   3   3   3   3   4   4   3   3   1   3   3   4   4   3   3   3   3   4
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   4   3   3   4   4   3   3   3   4   4   4   2   3   1   4   3   3   3   3   3
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   3   3   3   4   3   3   3   4   3   3   4   4   3   3   4   3   3   4   3   4
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   4   4   3   3   4   4   3   4   3   3   4   2   4   4   4   3   3   4   3   4
## 102 103 104 105 106 107 108 109
##   3   3   4   3   4   3   4   4
```

```
k.means$totss
```

```
## [1] 324
```

```
k.means$betweenss
```

```
## [1] 277.9691
```

```
k.means$withinss
```

```
## [1]  2.030908  1.211246 22.196961 20.591835
```

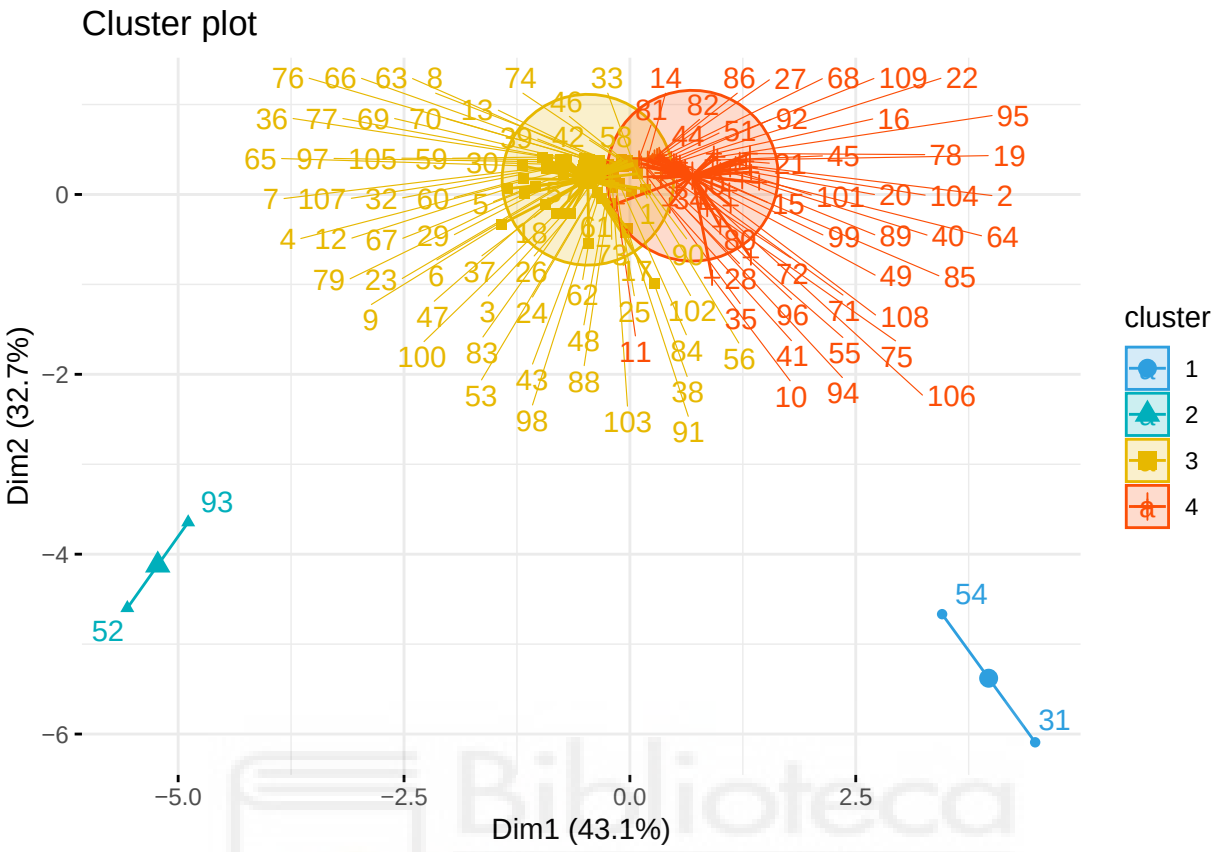
```
k.means$tot.withinss
```

```
## [1] 46.03095
```

Visualización de los clusters de k-means = 4

```
fviz_cluster(k.means, data = Z,
palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
ellipse.type = "euclid", # Concentration ellipse
star.plot = TRUE, # Add segments from centroids to items
repel = TRUE, # Avoid label overplotting (slow)
ggtheme = theme_minimal()
)
```

```
## Too few points to calculate an ellipse
## Too few points to calculate an ellipse
```



Utilizando K=4, se forman 4 clúster, vemos que el clúster 2 siguen lo forman Hong Kong y Singapur, el clúster 1 son China y India.

Conclusiones: Siempre me voy a quedar con la mayor suma de cuadrados entre grupos para saber como de bueno es el análisis. En K=2 se ha obtenido un valor de 107.53, en k=3 183.68 y en k=4 277.97, por lo tanto, vemos que la mejor opción es dividir a los individuos en 4 clúster (k=4).

5.2. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Para tipificar las variables - Analizar - Estadísticos descriptivos - Descriptivos - Pasamos las variables - Marcamos la opción de Guardar valores estandarizados como variables - Aceptar

Para realizar un análisis clúster no jerárquico - Analizar - clasificar - Clúster K-medias - pasamos las variables tipificadas, que nos aparecerán al final - En opciones marcamos la tabla ANOVA - En Gráficos marcamos Dendograma - Aceptar

K=2

Centros de clústeres iniciales

	Clúster 1	Clúster 2
Puntuación Z: Población x1000	-,28573	7,88867
Puntuación Z: Habitantes por Km2	7,82972	-,11753
Puntuación Z(urbana) Habitantes en ciudades (%)	1,54822	-1,26130

Historial de iteraciones[a]

Iteración	Cambiar en centros de clústeres 1	2
1	7,734	7,650
2	,127	,181
3	,014	,020
4	,015	,020
5	,014	,020
6	,015	,022
7	,000	,000

a. Convergencia conseguida debido a que no hay ningún cambio en los centros de clústeres o un cambio pequeño. El cambio de la coordenada máxima absoluta para

Centros de clústeres finales

	Clúster 1	2
Puntuación Z: Población x1000	-,12729	,18386
Puntuación Z: Habitantes por Km2	,09687	-,13327
Puntuación Z(urbana) Habitantes en ciudades (%)	,71271	-,99780

ANOVA

	Clúster Media cuadrática	gl	Error Media cuadrática	gl	F	Sig.
Puntuación Z: Población x1000	2,541	1	,994	106	2,556	,113
Puntuación Z: Habitantes por Km2	1,390	1	1,006	106	1,383	,242
Puntuación Z(urbana) Habitantes en ciudades (%)	76,803	1	,285	106	269,601	<,001

Las pruebas F sólo se deben utilizar con fines descriptivos porque los clústeres se han elegido para maximizar las diferencias entre los casos de distintos clústeres. Los niveles de significación observados no están corregidos para esto y, por lo tanto, no se pueden interpretar como pruebas de la hipótesis de que los medias de clúster son iguales.

K=3

Centros de clústeres iniciales

	Clúster		
	1	2	3
Puntuación Z: Población x1000	7,88867	-,30754	-,28573
Puntuación Z: Habitantes por Km2	-,11753	-,25072	7,82972
Puntuación Z(urbana) Habitantes en ciudades (%)	-1,26130	-,14576	1,54822

Historial de iteraciones[a]

	Cambiar en centros de clústeres		
Iteración	1	2	3
1	1,007	,257	,778
2	,000	,000	,000

a. Convergencia conseguida debido a que no hay ningún cambio en los centros de clústeres o un cambio pequeño. El cambio de la coordenada máxima absoluta para cualquier centro es ,000. La iteración actual es 2. La distancia mínimo entre los centros iniciales es 8,256.

Centros de clústeres finales

	Clúster		
	1	2	3
Puntuación Z: Población x1000	6,88817	-,12433	-,29561
Puntuación Z: Habitantes por Km2	,00013	-,13479	7,06164
Puntuación Z(urbana) Habitantes en ciudades (%)	-1,26130	-,00790	1,67217

ANOVA

	Clúster		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntuación Z: Población x1000	48,338	2	,107	105	450,792	<,001
Puntuación Z: Habitantes por Km2	50,811	2	,061	105	838,083	<,001
Puntuación Z(urbana) Habitantes en ciudades (%)	4,390	2	,935	105	4,693	,011

Las pruebas F sólo se deben utilizar con fines descriptivos porque los clústeres se han elegido para maximizar las diferencias entre los casos de distintos clústeres. Los niveles de significación observados no están corregidos para esto y, por lo tanto, no se pueden interpretar como pruebas de la hipótesis de que los medias de clúster son iguales.

K=4

Centros de clústeres iniciales

	Clúster			
	1	2	3	4
Puntuación Z: Población x1000	-,28437	7,88867	-,28573	-,31299
Puntuación Z: Habitantes por Km2	,01863	-,11753	7,82972	-,15749
Puntuación Z(urbana) Habitantes en ciudades (%)	-2,12895	-1,26130	1,54822	1,63086

Historial de iteraciones[a]

	Cambiar en centros de clústeres			
Iteración	1	2	3	4
1	1,129	1,007	,778	,996
2	,038	,000	,000	,029
3	,000	,000	,000	,000

a. Convergencia conseguida debido a que no hay ningún cambio en los centros de clústeres o un cambio pequeño. El cambio de la coordenada máxima absoluta para cualquier centro es ,000. La iteración actual es 3. La distancia mínimo entre los centros iniciales es 3,764.

Centros de clústeres finales

	Clúster			
	1	2	3	4
Puntuación Z: Población x1000	-,12797	6,88817	-,29561	-,12177
Puntuación Z: Habitantes por Km2	-,13948	,00013	7,06164	-,13148
Puntuación Z(urbana) Habitantes en ciudades (%)	-,98554	-1,26130	1,67217	,68125

ANOVA

	Clúster		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntuación Z: Población x1000	32,226	3	,108	104	297,694	<,001
Puntuación Z: Habitantes por Km2	33,875	3	,061	104	553,550	<,001
Puntuación Z(urbana) Habitantes en ciudades (%)	26,283	3	,271	104	97,103	<,001

Las pruebas F sólo se deben utilizar con fines descriptivos porque los clústeres se han elegido para maximizar las diferencias entre los casos de distintos clústeres. Los niveles de significación observados no están corregidos para esto y, por lo tanto, no se pueden interpretar como pruebas de la hipótesis de que los medias de clúster son iguales.

En las salidas de SPSS, en la tabla ANOVA nos fijamos en que la F de Snedecor tiene que ser mayor que 1. Siempre me voy a quedar con la suma de la Fs para saber como de bueno es el análisis.

Con k=2 obtenemos que la suma de las Fs es 273.54, con k=3 1293.568 y con k=4 948.35, por lo tanto, nos quedamos con 3 clústers, ya que la suma de las Fs es mayor, por lo que esta es la mejor opción de dividir a los individuos.

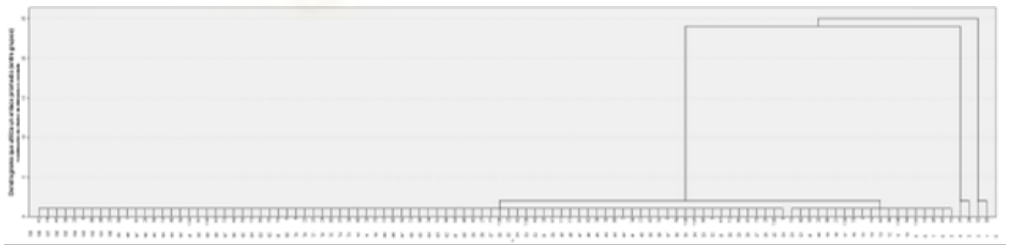


Figure 35: "DENDOGRAMA K=3"

En el dendograma se observan 3 grupos, uno muy grande y otros dos muy pequeños (al final).

5.3. Ejercicio 9. Repite el ejercicio anterior añadiendo la variable 'Esperanza de vida femenina' como variable de interés y resolviendo la clasificación con un análisis sí jerárquico.

Leemos los datos

```
mundo <- as.data.frame(read_sav("mundo.sav"))
```

Seleccionamos las variables de la base de datos del ejercicio anterior añadiendo la variable 'Esperanza de vida femenina':

```
datos <- subset(mundo, select = c("poblac", "urbana", "densidad","espvidaf"))
summary(datos)
```

```
##      poblac            urbana          densidad         espvidaf
##  Min.   :    256   Min.   :  5.00   Min.   :   2.3   Min.   :43.00
##  1st Qu.:   5100   1st Qu.: 40.75   1st Qu.:  29.0   1st Qu.:67.00
##  Median :  10400   Median : 60.00   Median :  64.0   Median :74.00
##  Mean   :  47724   Mean   : 56.53   Mean   : 203.4   Mean   :70.16
##  3rd Qu.:  35600   3rd Qu.: 75.00   3rd Qu.: 126.0   3rd Qu.:78.00
##  Max.   :1205200   Max.   :100.00   Max.   :5494.0   Max.   :82.00
##                    NA's   :1
```

Observamos si existen valores perdidos

```
any(is.na(datos))
```

```
## [1] TRUE
```

"TRUE" lo que nos indica que sí que hay valores perdidos.

Eliminamos las filas con valores faltantes

```
datos <- na.omit(datos) #eliminamos
```

Tipificamos los datos

```
G=cov.wt(datos, method='ML')$center
V=cov.wt(datos, method='ML')$cov
escala=diag(cov.wt(datos, method='ML')$cov)
Z=scale(datos,center=G,scale=sqrt(escala))
Z=as.matrix(Z)
```

El agrupamiento jerárquico es un análisis de agrupamiento sobre un conjunto de diferencias y métodos para analizarlo. Dicho agrupamiento se realiza mediante el uso de la hclust().

El argumento especifica una estructura de disimilitud producida por la dist() función. El segundo argumento es el method que especifica el método de aglomeración que se utilizará. Para realizar el Dendograma existen diferentes métodos, pero nosotros vamos a probar con los siguientes: average, complete y single.

```
hc1=hclust(dist(Z,method = "euclidean"),method="average") #la media
hc1 #calcula la media #dendogramas más uniformes
```

```
##
## Call:
## hclust(d = dist(Z, method = "euclidean"), method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 108
```

```
hc2=hclust(dist(Z,method = "euclidean"),method="complete")
hc2 #distancia entre dos grupos, el máximo entre los individuos de los grupos
```

```
##
## Call:
## hclust(d = dist(Z, method = "euclidean"), method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 108
```

```
hc3=hclust(dist(Z,method = "euclidean"),method="single")
hc3 #distancia entre dos grupos, el mínimo entre los individuos de los grupos
```

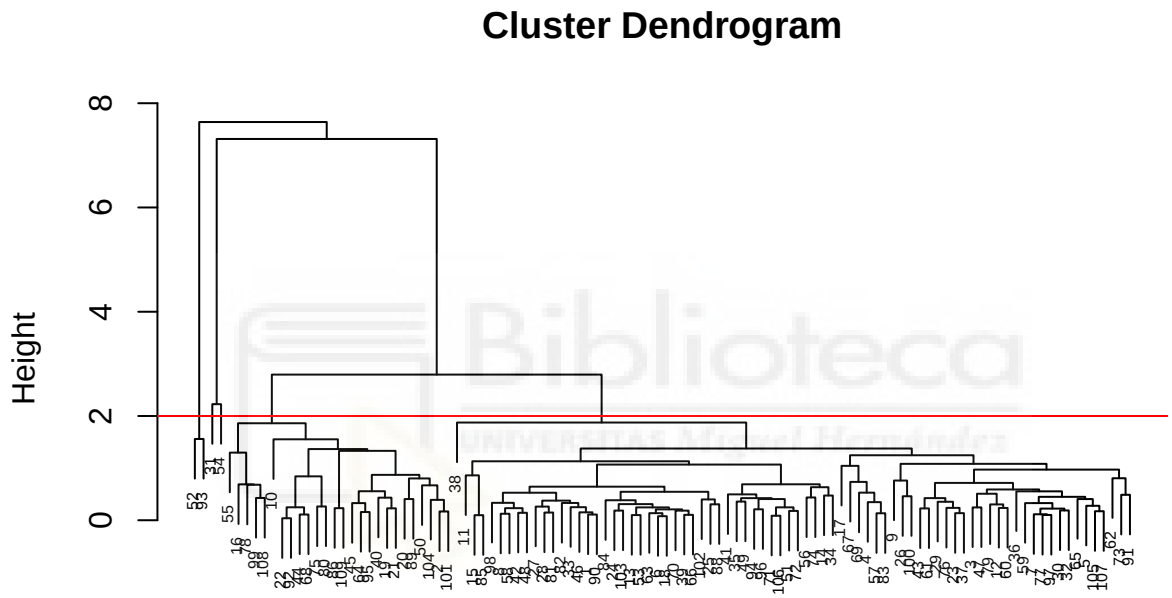
```
##
## Call:
## hclust(d = dist(Z, method = "euclidean"), method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 108
```

Si utilizamos el MÉTODO AVERAGE:

```
#calcula la media #dendogramas más uniformes

#Cluster jerarquico (dendrogram)
hc=hclust(dist(Z,method = "euclidean"),method="average")
hc
```

```
##
## Call:
## hclust(d = dist(Z, method = "euclidean"), method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 108
```

```
#Representamos el dendograma
plot(hc,cex="0.5")
abline(a=2,b=0,col="red")
```

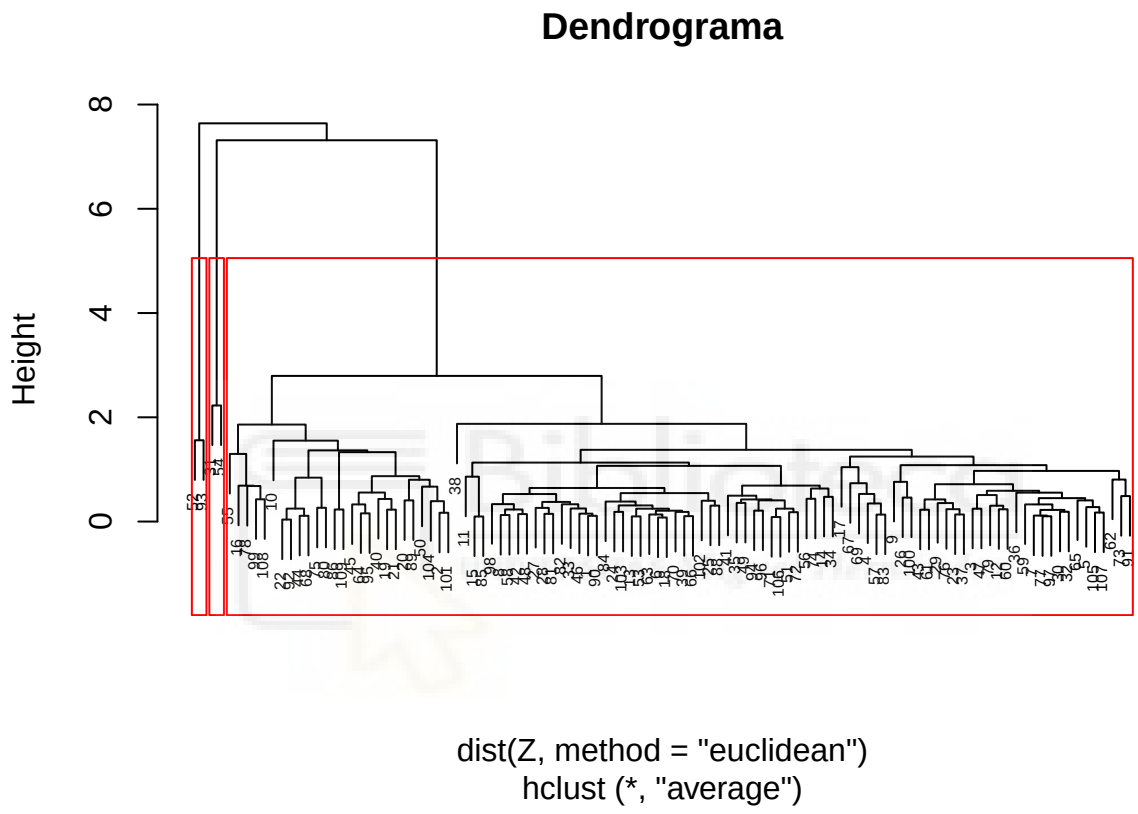



```
#partimos en grupos
grupos=cutree(hc,k=3)
grupos
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   1   3   1   2   1   1   1   1   1   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
```

```
##   1   1   1   1   1   1   1   1   1   1   1   3   1   1   1   1   1   1   1   1
## 102 103 104 105 106 107 108 109
##   1   1   1   1   1   1   1   1
```

```
#visualizamos el dendograma y los grupos:
plot(hc, main="Dendrograma", cex=0.5)
rect.hclust(hc, k=3, border="red")
```

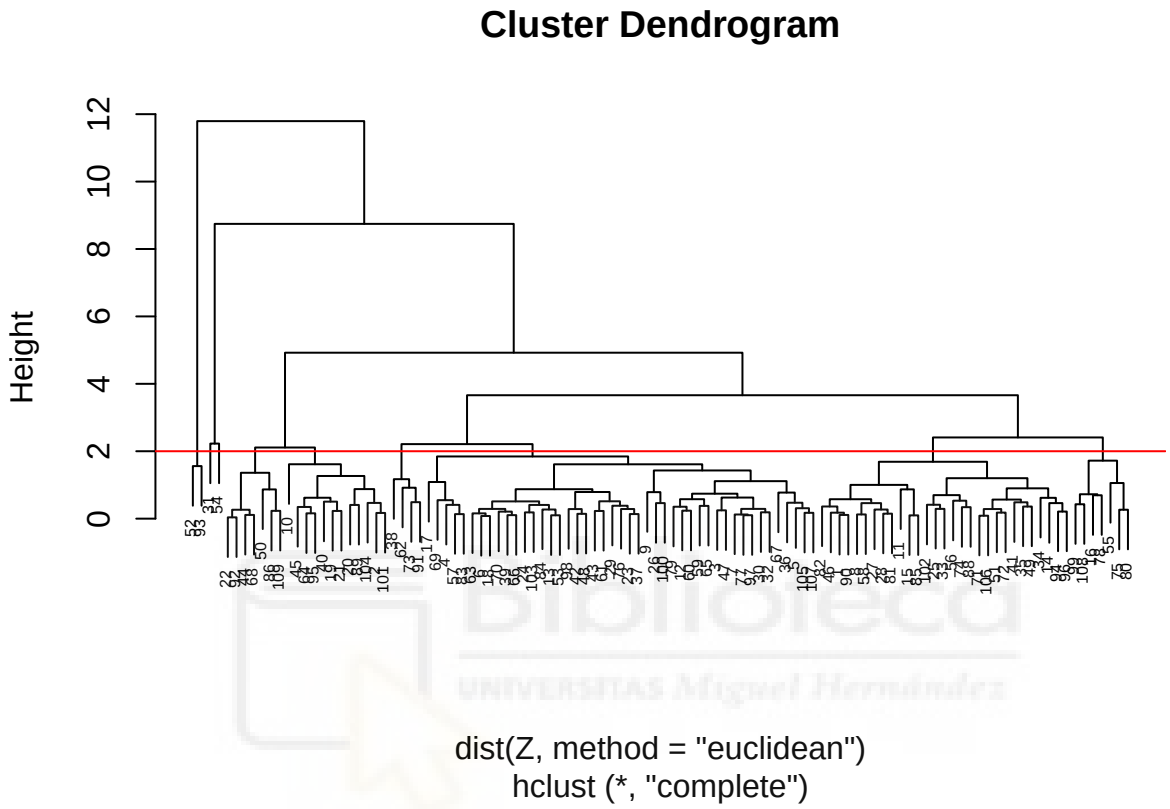


Si utilizamos el MÉTODO COMPLETE:

```
#distancia entre dos grupos, el máximo entre los individuos de los grupos

#Cluster jerarquico (dendrogram)
hc=hclust(dist(Z,method = "euclidean"),method="complete")
hc
```

```
##
## Call:
## hclust(d = dist(Z, method = "euclidean"), method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 108
```

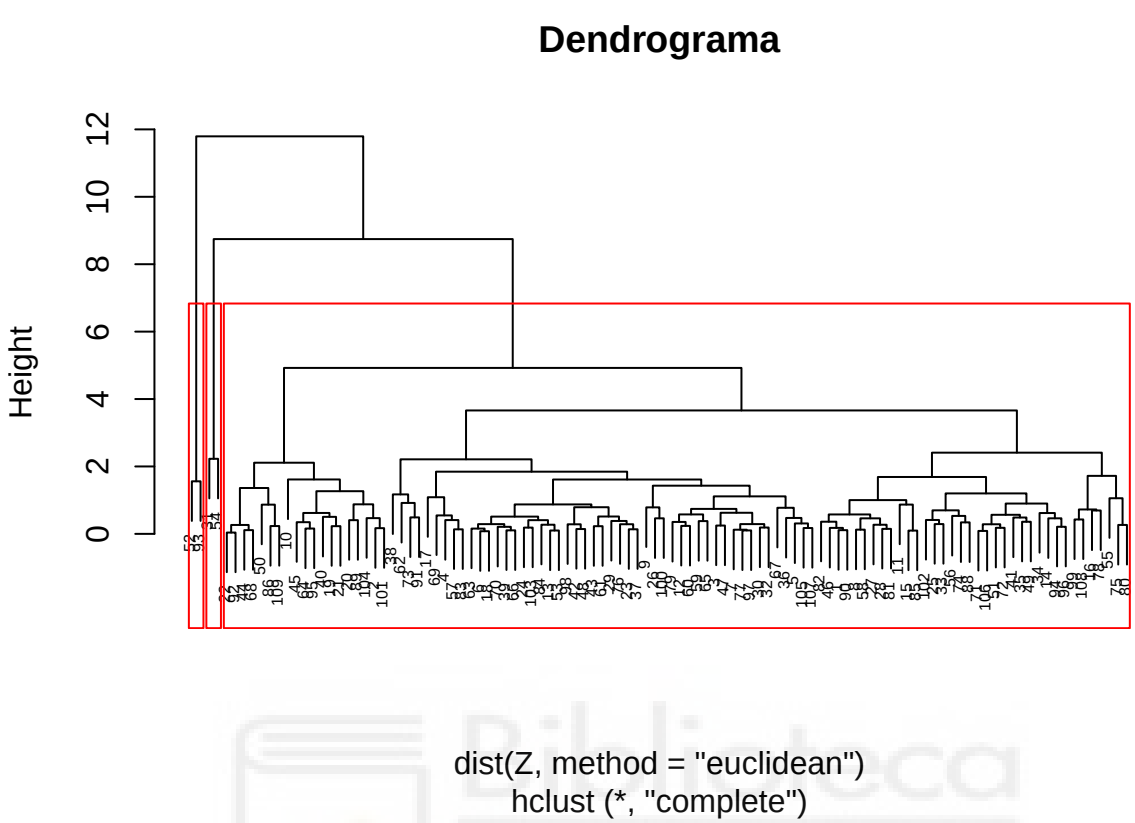
```
#Representamos el dendograma
plot(hc,cex="0.5")
abline(a=2,b=0,col="red")
```




```
#partimos en grupos
grupos=cutree(hc,k=3)
grupos
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   1   3   1   2   1   1   1   1   1   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   1   1   1   1   1   1   1   1   1   1   1   3   1   1   1   1   1   1   1   1
## 102 103 104 105 106 107 108 109
##   1   1   1   1   1   1   1   1
```

```
#visualizamos el dendograma y los grupos:
plot(hc, main="Dendrograma", cex=0.5)
rect.hclust(hc, k=3, border="red")
```

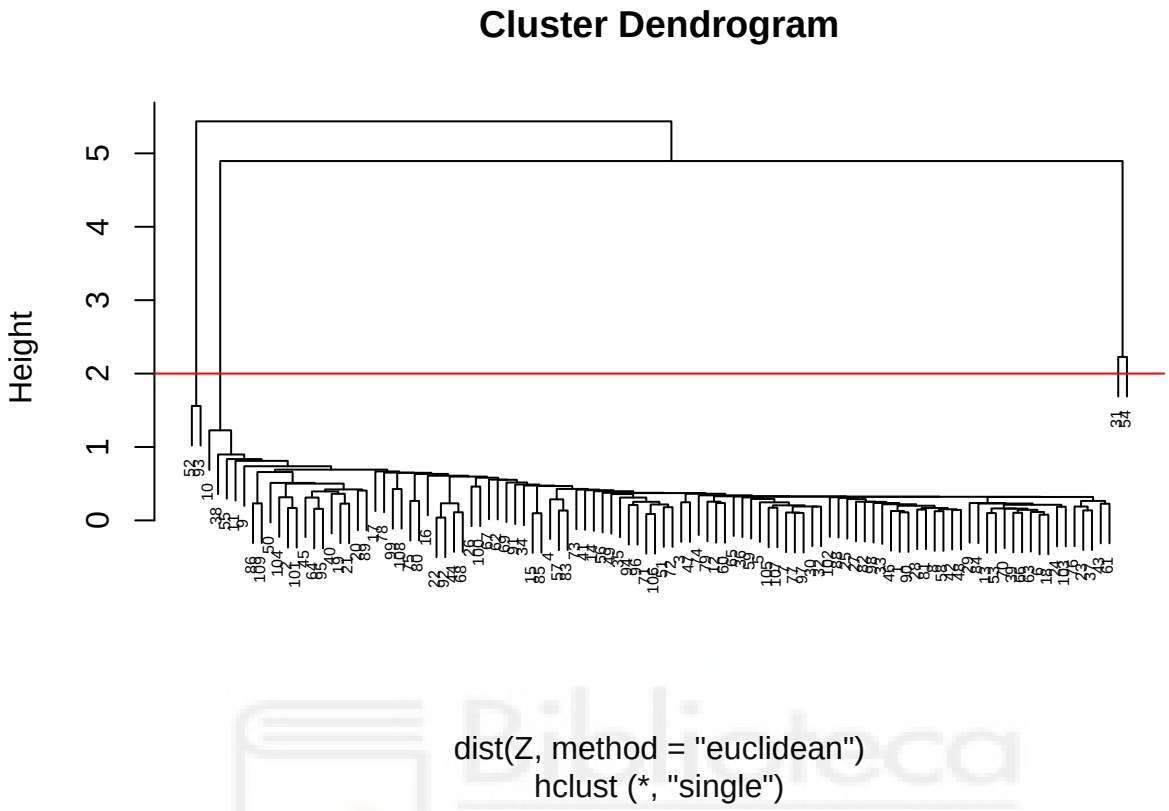


Si utilizamos el MÉTODO SINGLE:

```
#distancia entre dos grupos, el mínimo entre los individuos de los grupos

#Cluster jerarquico (dendrogram)
hc=hclust(dist(Z,method = "euclidean"),method="single")
hc
```

```
##
## Call:
## hclust(d = dist(Z, method = "euclidean"), method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 108
```

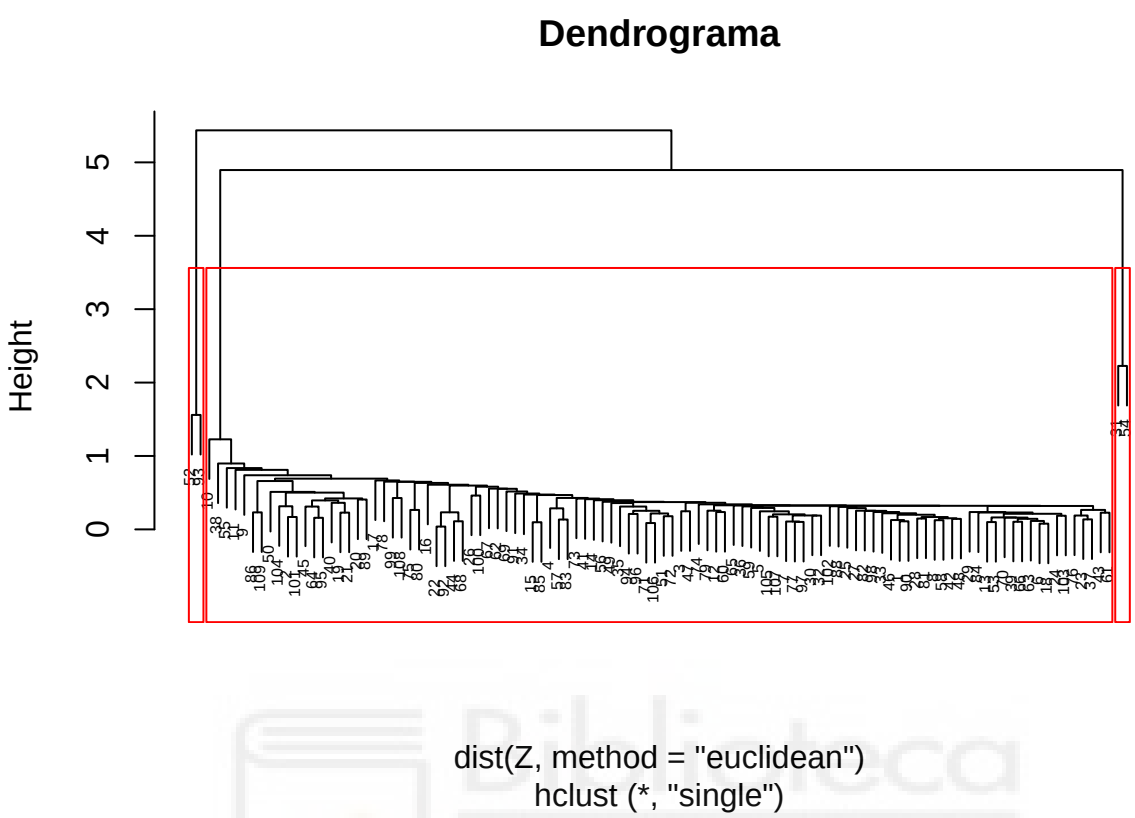
```
#Representamos el dendograma
plot(hc,cex="0.5")
abline(a=2,b=0,col="red")
```




```
#partimos en grupos
grupos=cutree(hc,k=3)
grupos
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   1   3   1   2   1   1   1   1   1   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   1   1   1   1   1   1   1   1   1   1   1   3   1   1   1   1   1   1   1   1
## 102 103 104 105 106 107 108 109
##   1   1   1   1   1   1   1   1
```

```
#visualizamos el dendograma y los grupos:
plot(hc, main="Dendrograma", cex=0.5)
rect.hclust(hc, k=3, border="red")
```



Independientemente del método que utilicemos se observan 3 grupos, uno muy grande y otros dos muy pequeños. Al igual que habíamos observado en el ejercicio anterior.

5.4. Los resultados obtenidos mediante R para este ejercicio se pueden validar mediante la salida en SPSS. Debemos seguir los siguientes pasos:

Tipificamos únicamente la variable nueva incorporada - Analizar - estadísticos descriptivos - descriptivos y pasamos solo 'espvidaf' - Marcamos la opción de Guardar valores estandarizados como variables - Aceptar

Analisis - cluster jerárquico y añadimos la variable 'espvidaf' tipificada - En gráficos marcamos dendograma

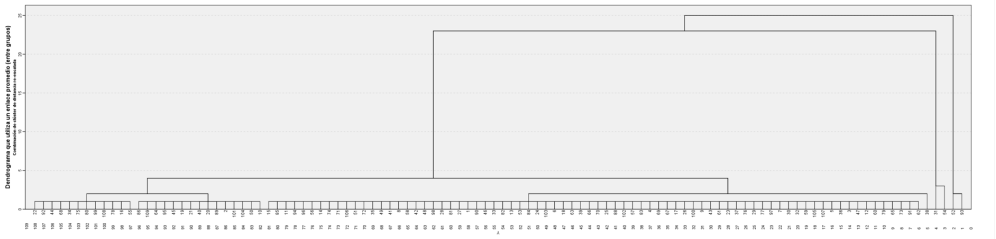


Figure 36: "DENDOGRAMA"

Igual que habíamos comentado en el ej anterior, se observan se observan 3 grupos, uno muy grande y otros dos muy pequeños (al final).

7. BIBLIOGRAFÍA

1. RPubs - Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/Joaquin_AR/287787

2. RPubs - Análisis de componentes principales (PCA) [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/Cristina_Gil/PCA

3. RPubs - Análisis Factorial [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/luis_bolanos/FA

4. RPubs - Análisis Factorial [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/luis_bolanos/FA

5. RPubs - Análisis de correspondencias simples [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/lucia_cabrera/610195

6. RPubs - Análisis de Correspondencia Múltiple (ACM) [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/ocamilocardona/813536

7. RPubs - Escalado multidimensional /análisis de correspondencia [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/luis_bolanos/CA-MDS

8. RPubs - Análisis de Cluster en R [Internet]. [citado 5 de junio de 2024]. Disponible en: https://rpubs.com/lhromeroj/analisisdeclusterR

9. Flores MJS. Capítulo 3 Análisis de Componentes Principales | Técnicas Multivariadas con R [Internet]. [citado 5 de junio de 2024]. Disponible en: https://bookdown.org/jsalinas/tecnicas_multivariadas/acp.html

10. Flores MJS. Capítulo 4 Análisis Factorial | Técnicas Multivariadas con R [Internet]. [citado 5 de junio de 2024]. Disponible en: https://bookdown.org/jsalinas/tecnicas_multivariadas/af.html

11. Flores MJS. Capítulo 6 Análisis Cluster | Técnicas Multivariadas con R [Internet]. [citado 5 de junio de 2024]. Disponible en: https://bookdown.org/jsalinas/tecnicas_multivariadas/cluster.html

12. Análisis Factorial en R - Diego Calvo [Internet]. [citado 5 de junio de 2024]. Disponible en: https://www.diegocalvo.es/analisis-factorial-en-r/

13. Escalamiento multidimensional en R - Diego Calvo [Internet]. [citado 5 de junio de 2024]. Disponible en: https://www.diegocalvo.es/escalamiento-multidimensional-en-r/

14. Escalamiento multidimensional con R | mtorr.dev | [Internet]. [citado 5 de junio de 2024]. Disponible en: https://mtorr.dev/blog/escalamiento_multidimensional/