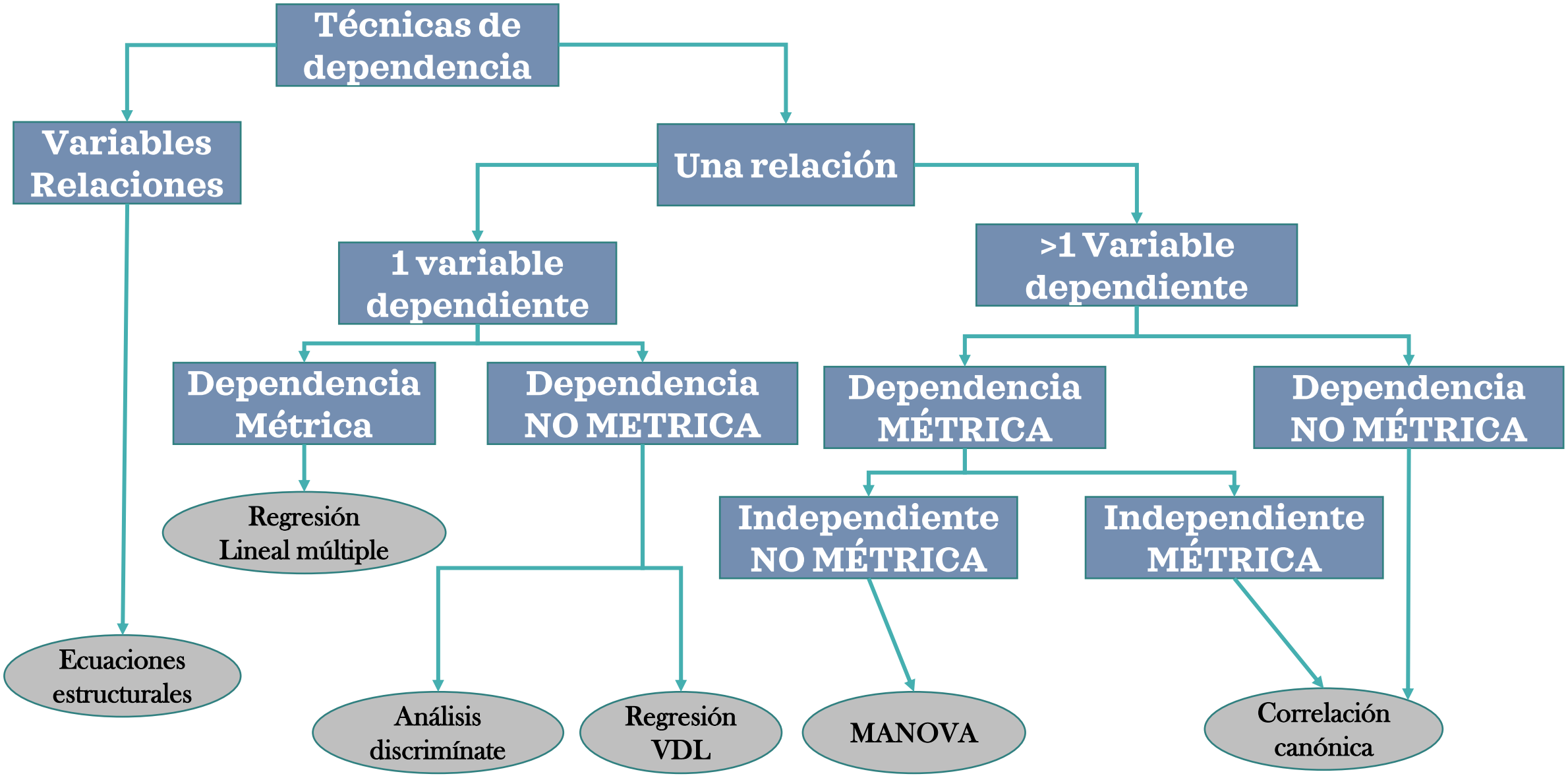
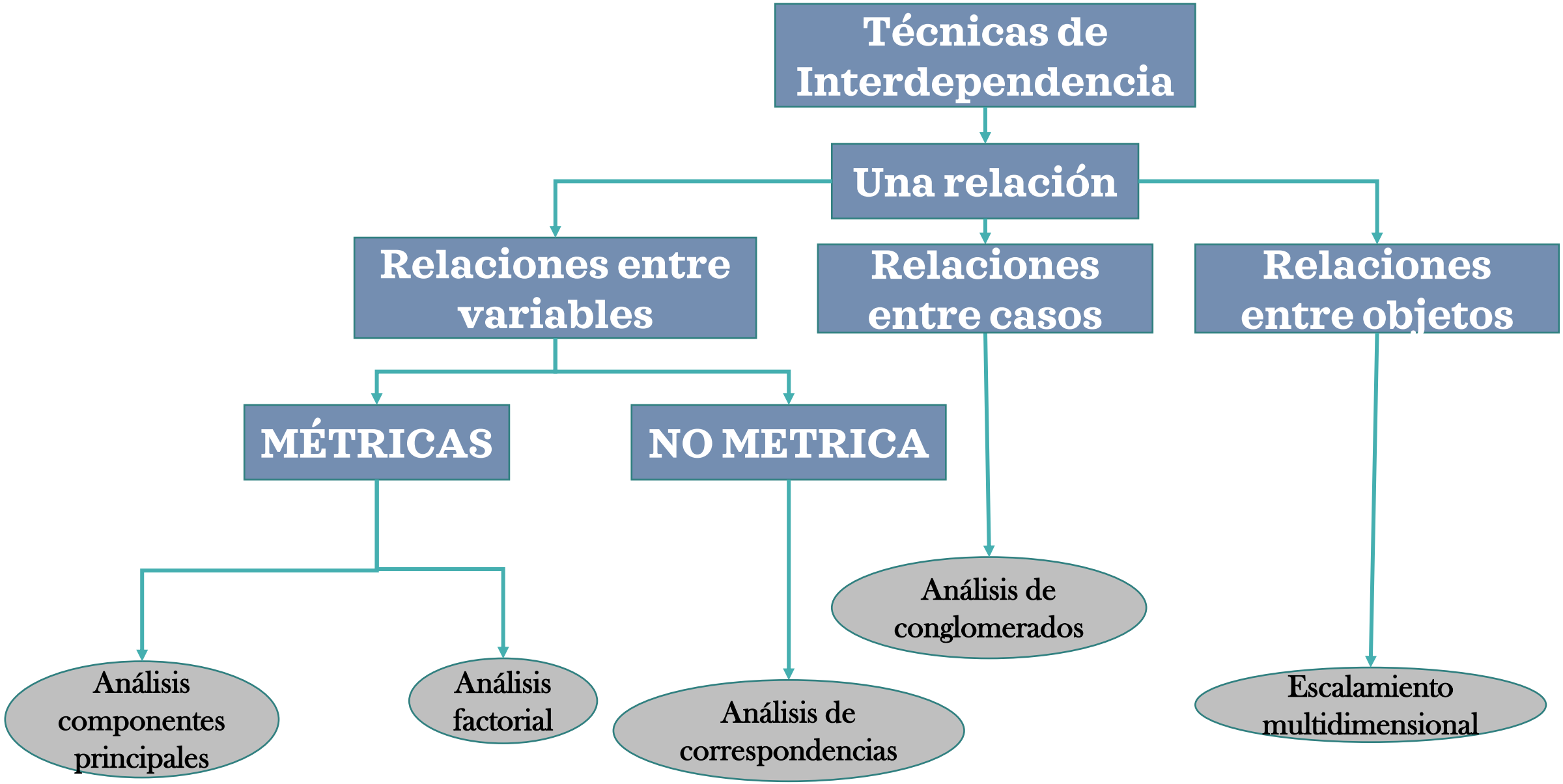
Análisis de Componentes Principales (ACP)

Dr. Misael Erikson Maguiña Palma

Técnica de análisis de dependencia



Técnica de análisis de dependencia



Fases de una investigación multivariante

- Definir el problema de Investigación
- Objetivos; diseño e hipótesis
- Técnica Multivariante a utilizar.

↓

- Desarrollo del Proyecto de Análisis

↓

- Evaluación de los supuestos de la Técnica Multivariante

↓

- Estimación del modelo Multivariante.
- Valoración del Ajuste del Modelo.

↓

- Interpretación de los valores Teóricos

↓

- Validación del Modelo Multivariante

Generalmente es mejor tener una muestra grande para validar el modelo. (2^(k) datos "k- variables")

Objetivo del ACP

El ACP es una técnica que proviene del análisis exploratorio de datos,

- Síntesis de la información, o reducción de la dimensión (numero de variables). Es decir reducir a u menor numero perdiendo la menor cantidad de información posible.

Ojo: las nuevas variables se llaman componentes.

- Uso solo para variables cuantitativas.

Ejemplo

id	var1	var2	...	varn
Id1				
Id2				
. . .				
Idm				

Var	Comp1	comp2
var1	+	
var2	+	
. . .		+
varn		+

Objetivo del ACP

- El objetivo es construir un pequeño numero de nuevas variables llamadas (componentes) en las cuales se concentra la mayor cantidad posible de información.

Tabla de datos

$$\begin{pmatrix} X_{11} & X_{21} & \cdots & X_{1p} \\ X_{12} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

100% de la información

Componentes

$$\begin{pmatrix} C_{11} & C_{21} & \cdots & C_{1p} \\ C_{12} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{np} \end{pmatrix}$$

80% **16%** **0.02%**

Permite ver los datos en R2, para identificar segmentos, clúster, similitudes y disimilitudes

Algunas consideraciones

- Cuando se recoge la información de una muestra de datos. Es usual tomar un gran numero de variables, lo cual dificulta visualizar las relaciones de las variables.
- Otro problema que se presenta es la fuerte correlación que muchas veces se presente entre las variables : si tomamos demasiadas variables, lo normal es que estén relacionados o que midan lo mismo bajo distintos puntos de vista.

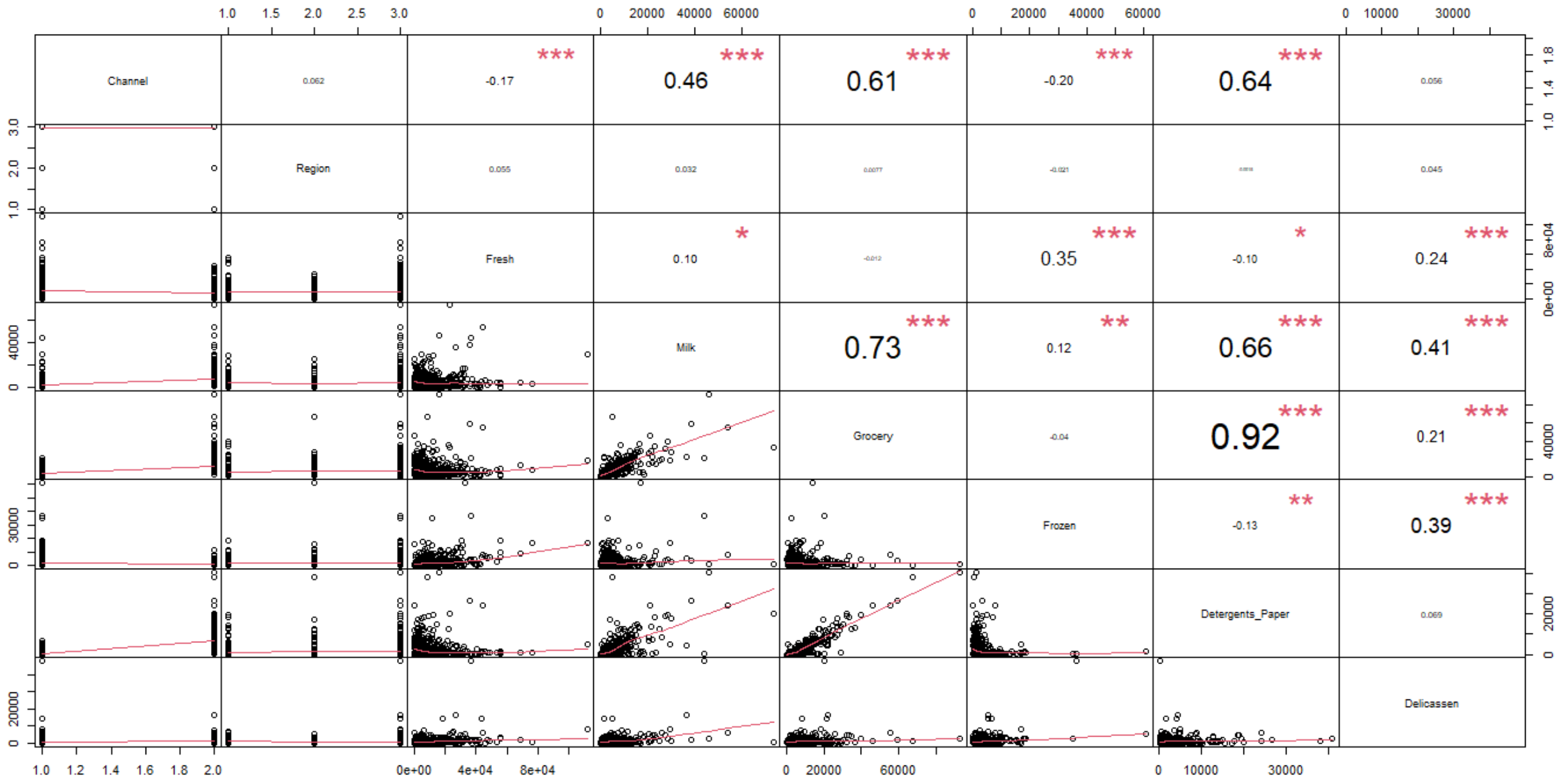
Fases de un análisis ACP

1.Diagnostico de Correlaciones

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Fresh	1.000	0.101	-0.012	0.346	-0.102	0.245
Milk	0.101	1.000	0.728	0.124	0.662	0.406
Grocery	-0.012	0.728	1.000	-0.040	0.925	0.205
Frozen	0.346	0.124	-0.040	1.000	-0.132	0.391
Detergents_Paper	-0.102	0.662	0.925	-0.132	1.000	0.069
Delicassen	0.245	0.406	0.205	0.391	0.069	1.000

Si las variables originales (x1, . . . , xp) están incorreladas, entonces carece de sentido calcular unos componentes principales

Matriz de correlaciones y significancia



Aplicación de test de correlación

La Prueba de esfericidad de Bartlett contrasta si la matriz de correlaciones es una matriz identidad, lo cual indicaría que el modelo factorial es inadecuado.

Test o Prueba de coeficiente de correlación

```
Tests of correlation matrices
Call:psych::cortest(R1 = cor(bacp))
 Chi Square value 597.98  with df =  28   with probability <
 3.6e-108
```

Test o Prueba Barlett

```
$chisq
[1] 1732.752

$p.value
[1] 0

$df
[1] 28
```

El estadístico de Bartlett se obtiene a partir de una transformación X^2 del determinante de la matriz de correlaciones y cuanto mayor sea, y por tanto menor el nivel de significación, más improbable es que la matriz sea una matriz identidad y más adecuado resulta el análisis factorial.

Kaiser-Meyer-Olkin (Coeficiente KMO)

Contrasta si las correlaciones parciales entre las variables son pequeñas, toma valores entre 0 y 1, e indica que el análisis factorial es tanto más adecuado cuanto mayor sea su valor. Así, Kaiser propuso en 1974 el siguiente criterio para decidir sobre la adecuación del análisis factorial de un conjunto de datos:

$0{,}9 \leq KMO \leq 1{,}0 \mapsto$ Excelente adecuación muestral

$0{,}8 \leq KMO \leq 0{,}9 \mapsto$ Buena adecuación muestral

$0{,}7 \leq KMO \leq 0{,}8 \mapsto$ Aceptable adecuación muestral

$0{,}6 \leq KMO \leq 0{,}7 \mapsto$ Regular adecuación muestral

$0{,}5 \leq KMO \leq 0{,}6 \mapsto$ Mala adecuación muestral

$0{,}0 \leq KMO \leq 0{,}5 \mapsto$ Adecuación muestral inaceptable

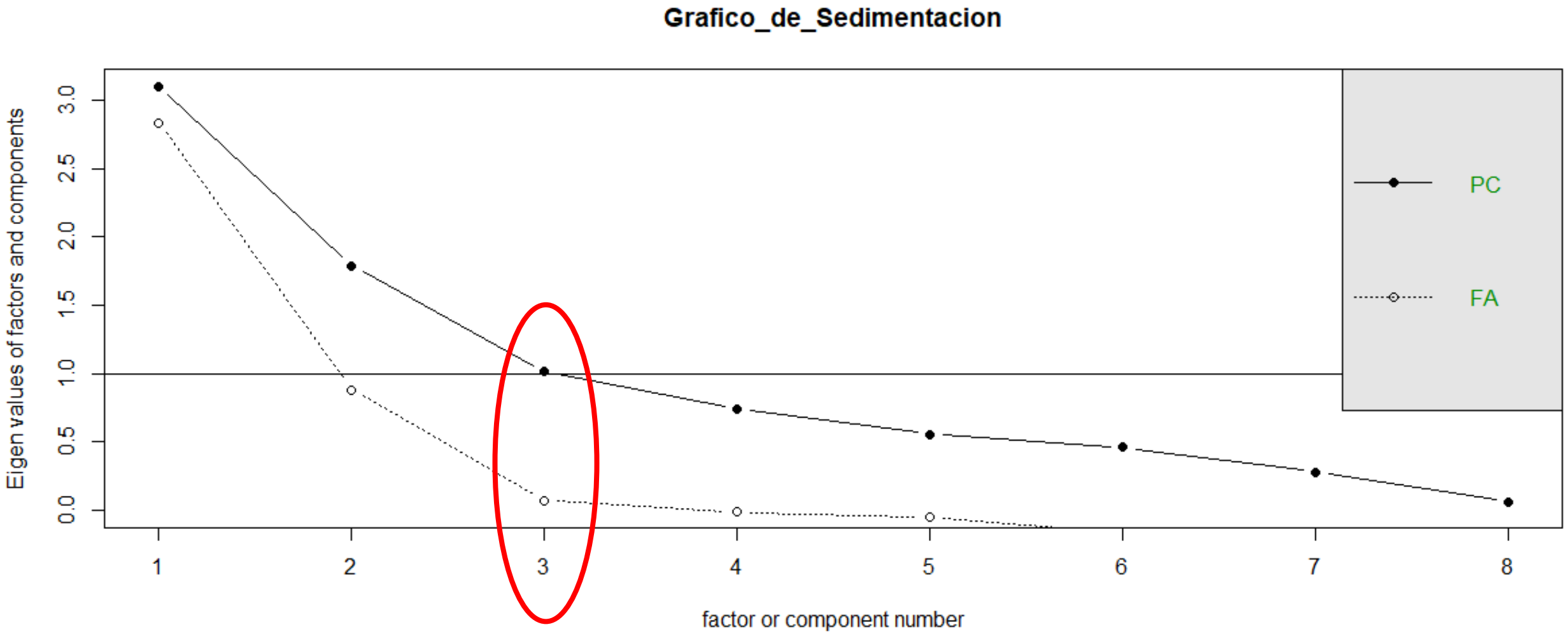
Kaiser-Meyer-Olkin (Coeficiente KMO)

```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = bacp)
Overall MSA =  0.73
MSA for each item =
          Channel             Region              Fresh
             0.92               0.41               0.68
             Milk            Grocery             Frozen
             0.87               0.69               0.67
Detergents_Paper         Delicassen
             0.68               0.59
```

Con el comando “KMO” podemos realizar el test de KMO a la data para lo cual el valor de Overall MSA debería ser mayor a 0.5 para justificar la realización de PCA en la data para valores menor esa 0.5 la realización del PCA no seria justificada

Número de componentes principales

Grafico de Sedimentación



Número de componentes principales

Criterio del porcentaje: El número m de componentes principales se toma de modo que Pm sea próximo a un valor especificado por el usuario, por ejemplo el 80%. Por otra parte, si la representación de P1, P2 ,…,Pk,… con respecto de k prácticamente se estabiliza a partir de un cierto m, entonces aumentar la dimensión apenas aporta más variabilidad explicada.

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.7607 1.3379 1.0059 0.8593 0.74608 0.67772 0.53021
Proportion of Variance 0.3875 0.2238 0.1265 0.0923 0.06958 0.05741 0.03514
Cumulative Proportion  0.3875 0.6112 0.7377 0.8300 0.89960 0.95701 0.99215
```

Número de componentes principales

Criterio de Kaiser: Obtener las componentes principales a partir de la matriz de correlaciones R equivale a suponer que las variables observables tengan varianza 1. Por lo que indica que hay que conservar los componentes principales cuyos valores propios son mayores que la unidad, aunque el criterio más utilizado es el de observar el porcentaje de varianza total explicada por cada componente o factor, y cuando éste llega a un porcentaje acumulado considerado alto, normalmente cerca del ochenta por ciento.

Desviación Estándar

```
1.7606845 1.3378965 1.0058697 0.8592976 0.7460780
0.6777229 0.5302132 0.2505796
```

varianza

<3

```
3.10000983 1.78996704 1.01177388 0.73839230 0.55663240
0.45930835 0.28112605 0.06279015
```

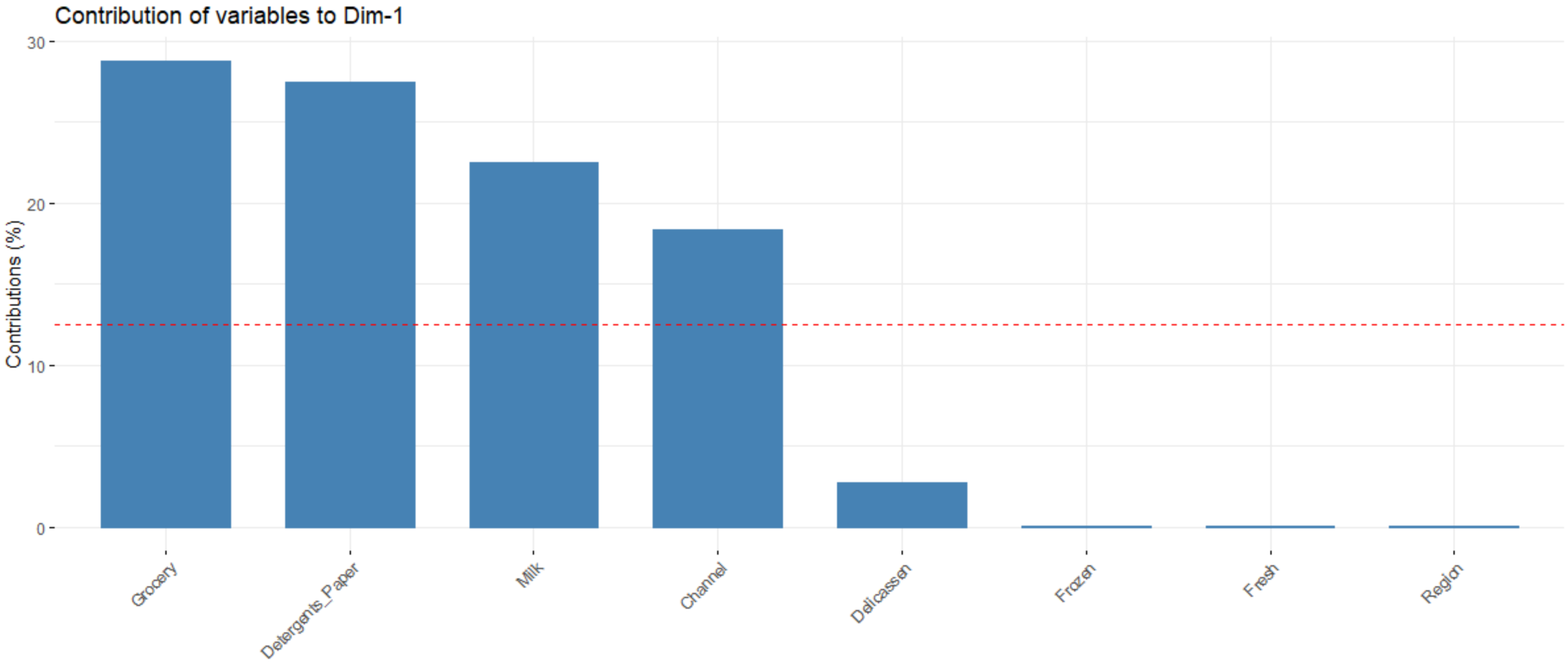
Realización de los Componentes

	PC1	PC2	PC3	PC4
Channel	-0.42829156	0.20469886	0.0829798863	-0.02964416
Region	-0.02472603	-0.04312964	0.9825008891	-0.07784462
Fresh	0.02531946	-0.51344468	0.0889509074	0.79847592
Milk	-0.47440995	-0.20554061	-0.0257510842	-0.05402202
Grocery	-0.53632914	0.00871762	-0.0453143572	0.12158624
Frozen	0.02997456	-0.59274525	-0.1221565222	-0.16131688
Detergents_Paper	-0.52390630	0.12108309	-0.0474814388	0.15101211
Delicassen	-0.16499653	-0.53318082	0.0009301994	-0.53755767

Notamos que hay aporte negativo de las variables: channel, milk, grocery, detergents, para el componente 1, mientras que para el componente 2 las variables que aportan son (fresh, frozen y delicassen=Productos fríos) y para el componente pc3 únicamente la región.

Productos fríos=-0.51*fresh -0.59*frozen -0.53*delicassen

Contribuciones de variables



Contribuciones de variables

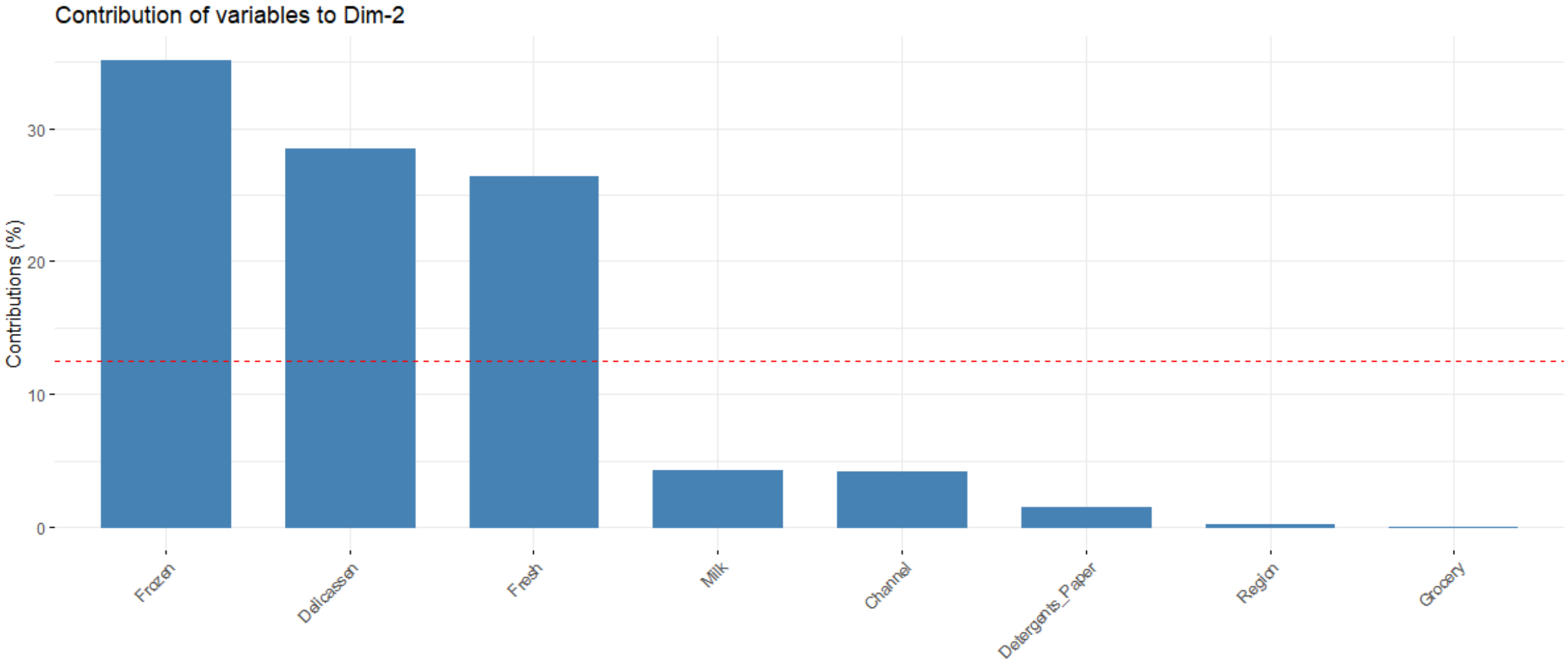


Grafico de la data en cada componente

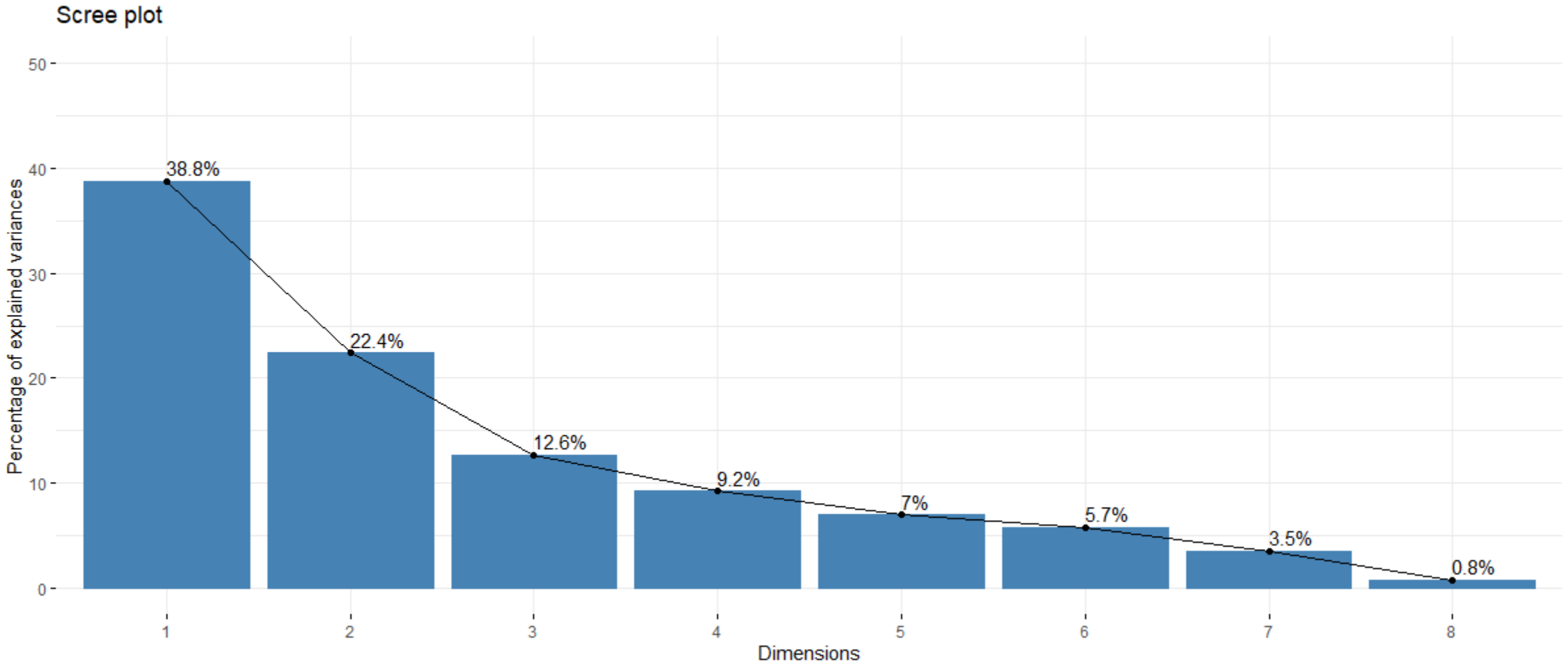


Gráfico de individuos.

Las personas con un perfil similar se agrupan.

Factores individuales

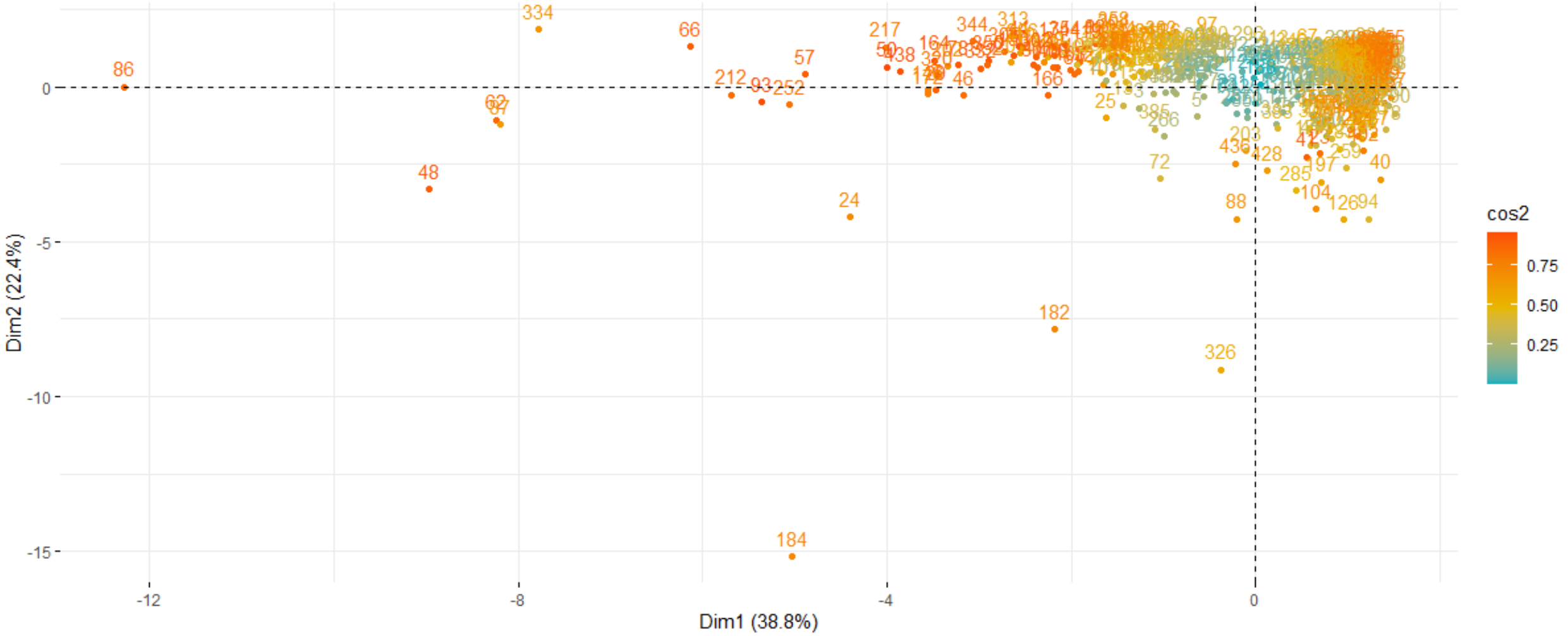
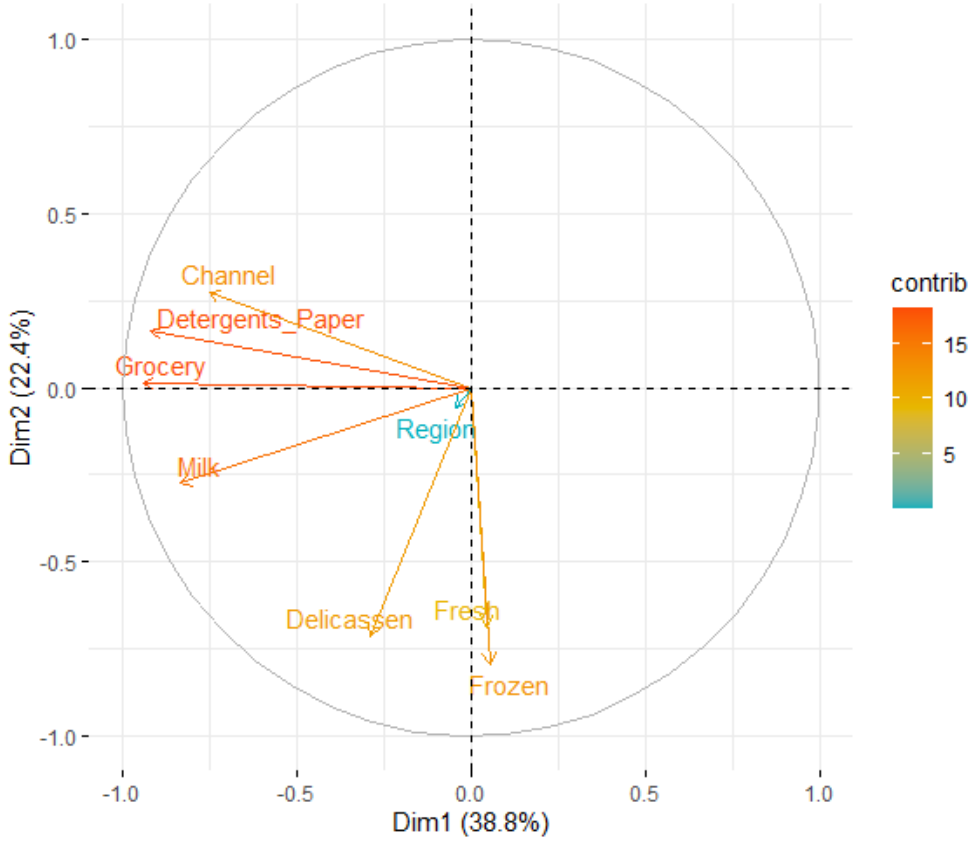


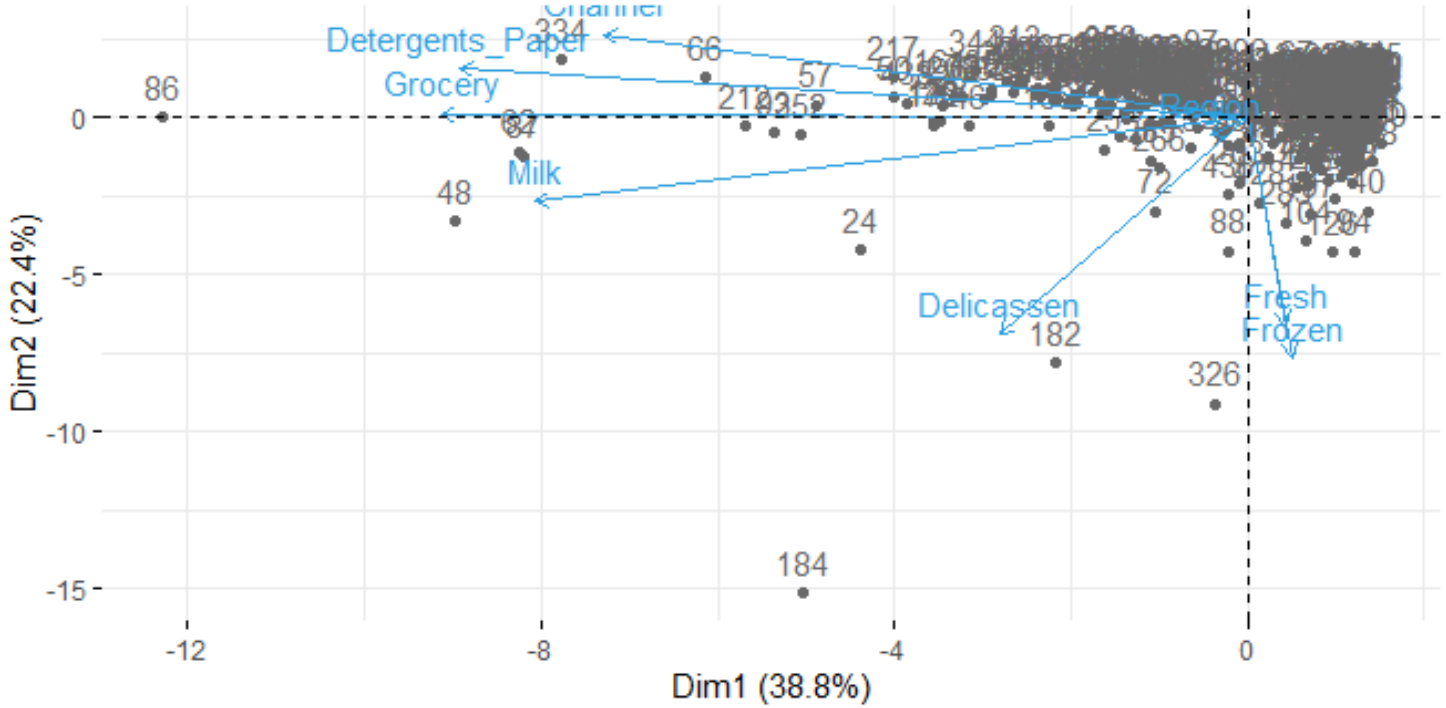
Gráfico de individuos.

Las personas con un perfil similar se agrupan.

Circulo de correlaciones

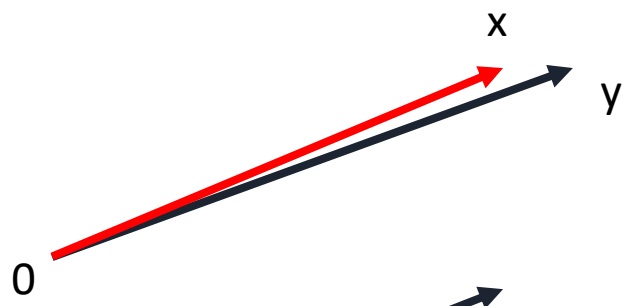


Biplot de individuos y variables



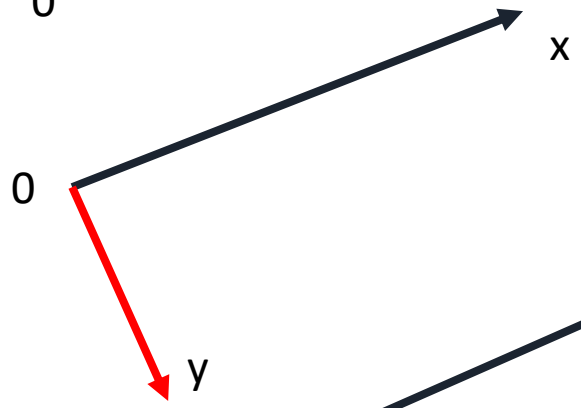
Correlaciones entre variables

CASO 1:



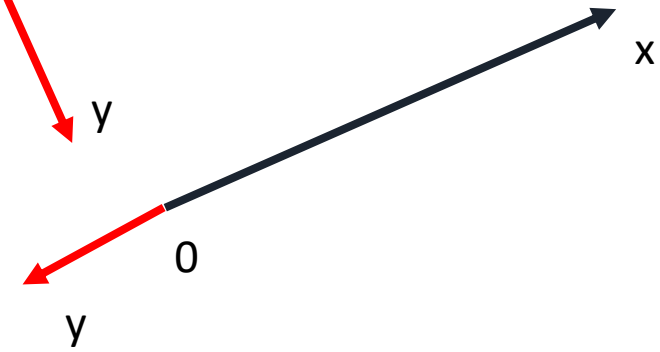
Θ=0° implica que Cos(Θ)=R(x,y)=1

CASO 2:



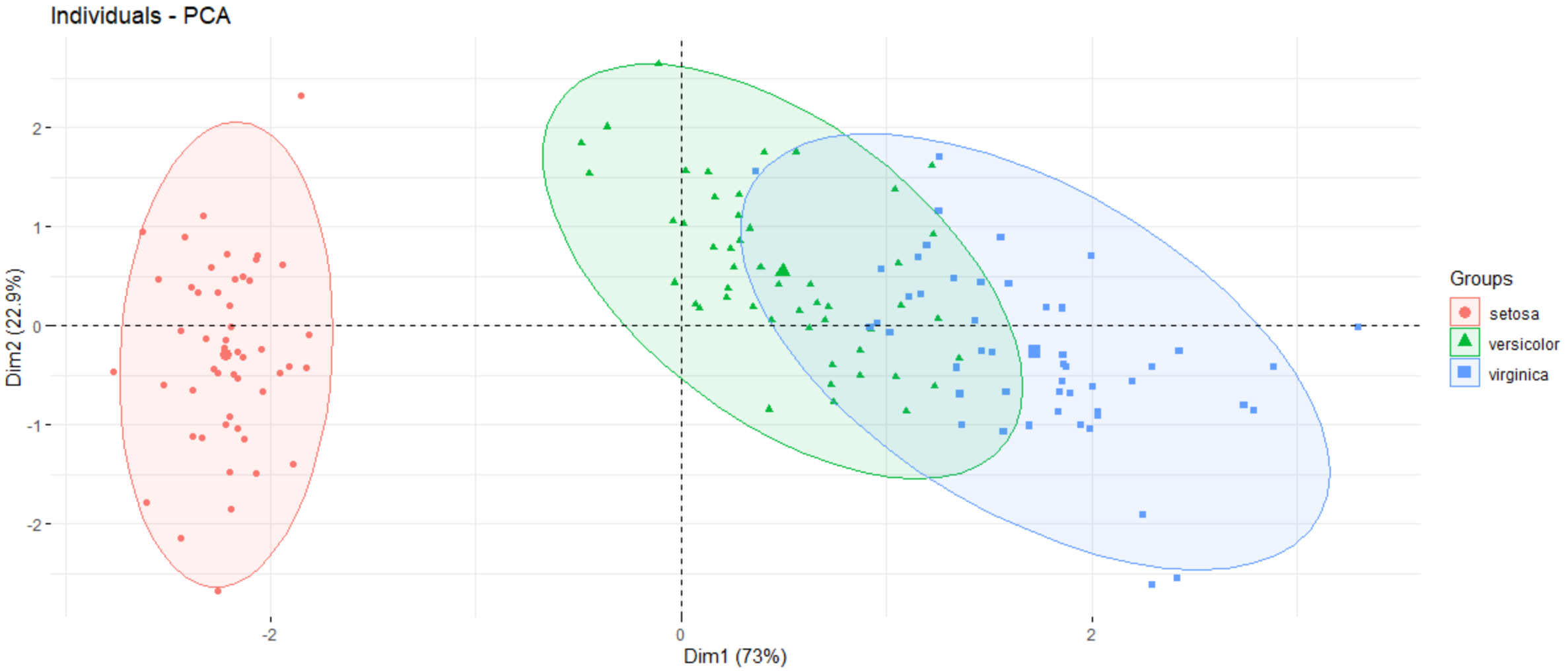
Θ=90° implica que Cos(Θ)=R(x,y)=0

CASO 3:



Θ=180° implica que Cos(Θ)=R(x,y)=-1

Variables cualitativas / categóricas



prcomp() and princomp() functions

El formato simplificado de estas 2 funciones es:

prcomp (x, escala = FALSO)

princomp (x, cor = FALSE, score = TRUE)

Argumentos para prcomp ():

x: una matriz numérica o marco de datos

scale: un valor lógico que indica si las variables deben escalarse para tener varianza unitaria antes de que se lleve a cabo el análisis

Argumentos para princomp ():

x: una matriz numérica o marco de datos

cor: un valor lógico. Si es TRUE, los datos se centrarán y escalarán antes del análisis.

scores: un valor lógico. Si es TRUE, se calculan las coordenadas de cada componente principal

Los elementos de las salidas devueltas por las funciones prcomp () y princomp () incluyen

Paquete para visualización PCA

install.packages("factoextra")

get_eigenvalue (res.pca): Extraiga los valores propios / varianzas de los componentes principales

fviz_eig (res.pca): Visualiza los valores propios

get_pca_ind (res.pca), get_pca_var (res.pca): Extrae los resultados para individuos y variables, respectivamente.

fviz_pca_ind (res.pca), fviz_pca_var (res.pca): Visualiza los resultados individuales y variables, respectivamente.

fviz_pca_biplot (res.pca): Haz una biplot de individuos y variables.