A2-Componentes Principales

Carlos David Lozano Sanguino-A01275316

2023-09-29

PARTE I

A partir de los datos sobre indicadores económicos y sociales de 96 países (datos: paises_mundo.csv Download paises_mundo.csv) hacer una análisis de Componentes principales a partir de la matriz de varianzas-covarianzas y otro a partir de la matriz de correlaciones , comparar los resultados y argumentar cuál es mejor según los resultados obtenidos.

```
library(readxl)
M = read_excel("C:/Users/CDLS1/Downloads/paises_mundo.xlsx")
head(M, n=5)
```

```
## # A tibble: 5 x 11
##   CrecPobl MortInf PorcMujeres  PNB95 ProdElec LinTelf ConsAgua PropBosq
##      <dbl>   <dbl>       <dbl>  <dbl>    <dbl>   <dbl>    <dbl>    <dbl>
## 1      1        30          41   2199     3903      12       94       53
## 2      3       124          46   4422      955       6       57       19
## 3      4.3      21          13 133540    91019      96      497        1
## 4      2.5      34          24  44609    19883      42      180        2
## 5      1.3      22          31 278431    65962     160     1043       22
## # i 3 more variables: PropDefor <dbl>, ConsEner <dbl>, EmisCO2 <dbl>
```

Paso 1: Calcular la matriz de varianzas-covarianzas (S) y la matriz de correlaciones (R).

```
# Matriz de varianzas-covarianzas
cov_matrix <- cov(M)

# Matriz de correlaciones
cor_matrix <- cor(M)
```

Paso 2: Calcular los valores y vectores propios de ambas matrices.

```
# Valores y vectores propios de la matriz de varianzas-covarianzas
eigen_cov <- eigen(cov_matrix)

# Valores y vectores propios de la matriz de correlaciones
eigen_cor <- eigen(cor_matrix)
```

Paso 3: Calcular la proporción de varianza explicada por cada componente principal.

```
# Proporción de varianza explicada por cada componente principal (matriz de varianzas-covarianzas)
var_exp_cov <- eigen_cov$values / sum(diag(cov_matrix))

# Proporción de varianza explicada por cada componente principal (matriz de correlaciones)
var_exp_cor <- eigen_cor$values / sum(diag(cor_matrix))
```

Paso 4: Acumular los resultados anteriores.

```
# Acumulación de la varianza explicada (matriz de varianzas-covarianzas)
cum_var_exp_cov <- cumsum(var_exp_cov)

# Acumulación de la varianza explicada (matriz de correlaciones)
cum_var_exp_cor <- cumsum(var_exp_cor)

# Imprimir resultados
print("Proporción de varianza explicada por cada componente (Matriz de varianzas-covarianzas):")
```

```
## [1] "Proporción de varianza explicada por cada componente (Matriz de varianzas-covarianzas):"
```

```
print(var_exp_cov)
```

```
##  [1] 9.034543e-01 9.647298e-02 6.795804e-05 4.554567e-06 1.782429e-07
##  [6] 7.530917e-09 5.317738e-09 6.657763e-10 8.502887e-11 2.107843e-11
## [11] 6.989035e-12
```

```
print("Varianza acumulada por cada componente (Matriz de varianzas-covarianzas):")
```

```
## [1] "Varianza acumulada por cada componente (Matriz de varianzas-covarianzas):"
```

```
print(cum_var_exp_cov)
```

```
##  [1] 0.9034543 0.9999273 0.9999953 0.9999998 1.0000000 1.0000000 1.0000000
##  [8] 1.0000000 1.0000000 1.0000000 1.0000000
```

```
print("Proporción de varianza explicada por cada componente (Matriz de Correlaciones):")
```

```
## [1] "Proporción de varianza explicada por cada componente (Matriz de Correlaciones):"
```

```
print(var_exp_cor)
```

```
##  [1] 0.366352638 0.175453813 0.124582832 0.078592361 0.072194597 0.066290906
##  [7] 0.051936828 0.029709178 0.015278951 0.013302563 0.006305332
```

```
print("Varianza acumulada por cada componente (Matriz de Correlaciones):")
```

```
## [1] "Varianza acumulada por cada componente (Matriz de Correlaciones):"
```

```
print(cum_var_exp_cor)
```

```
##  [1] 0.3663526 0.5418065 0.6663893 0.7449816 0.8171762 0.8834671 0.9354040
##  [8] 0.9651132 0.9803921 0.9936947 1.0000000
```

Paso 5 y 6. Analisis de Componentes Principales para ambas Matrices:

Matriz de Varianzas-Covarianzas:

- La primera componente principal explica aproximadamente el 90.35% de la varianza total de los datos.
- La segunda componente principal explica alrededor del 9.65% de la varianza total.
- Las componentes restantes explican cantidades muy pequeñas de varianza, con valores cercanos a cero.
- Las primeras dos componentes principales capturan la mayoría de la varianza en los datos, lo que sugiere que son las más significativas

Matriz de Correlaciones:

- La primera componente principal explica aproximadamente el 36.64% de la varianza total de los datos.
- La segunda componente principal explica alrededor del 17.55% de la varianza total.
- A diferencia de la matriz de varianzas-covarianzas, la matriz de correlaciones muestra una distribución de varianza explicada más uniforme entre las primeras dos componentes principales.
- Las primeras dos componentes principales siguen siendo las más importantes, pero la diferencia en la varianza explicada entre ellas es más equitativa.

Paso 7 Compare los resultados de los incisos 5y 6. ¿qué concluye?

Comparando los resultados de ambos enfoques (matriz de varianzas-covarianzas vs. matriz de correlaciones):

En la matriz de varianzas-covarianzas, las dos primeras componentes principales explican conjuntamente aproximadamente el 99.99% de la varianza total. Esto sugiere que estas dos componentes son altamente informativas y pueden utilizarse para reducir la dimensionalidad de los datos sin perder mucha información.

En la matriz de correlaciones, aunque las dos primeras componentes principales también son importantes, la varianza se distribuye de manera más uniforme entre ellas. Esto podría ser útil si estás interesado en mantener una representación más equilibrada de las variables originales.

PARTE II

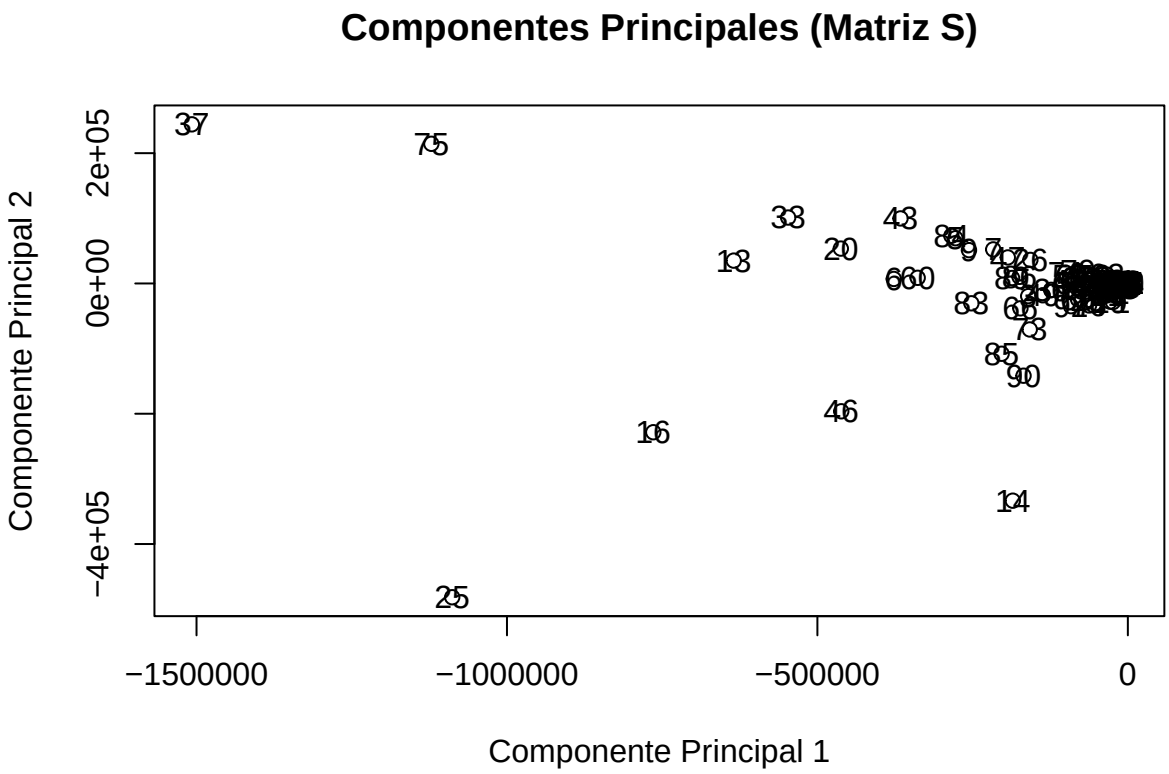
Obtenga las gráficas de respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes e interprete los resultados en término de agrupación de variables (puede ayudar “índice de riqueza”, “índice de ruralidad”)

```
library(stats)
library(factoextra)
```

```
## Loading required package: ggplot2
```

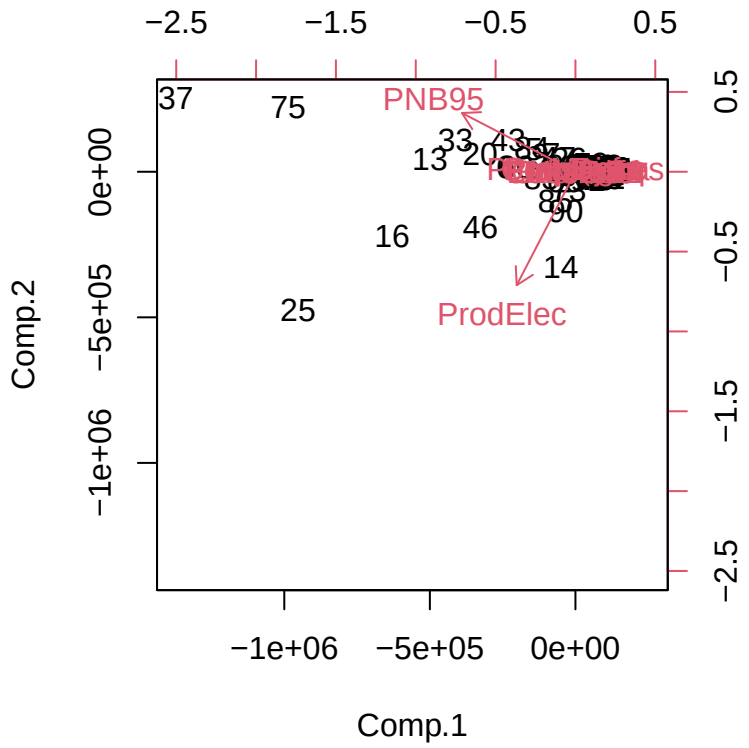
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ggplot2)
datos=M
# Proyectar los datos originales en las dos primeras componentes principales (S)
cpS <- princomp(datos, cor = FALSE)
cpaS <- as.matrix(datos) %*% cpS$loadings

# Gráfica de dispersión de las dos primeras componentes principales (S)
plot(cpaS[, 1], cpaS[, 2], type = "p", main = "Componentes Principales (Matriz S)",
     xlab = "Componente Principal 1", ylab = "Componente Principal 2")
text(cpaS[, 1], cpaS[, 2], labels = row.names(datos))
```

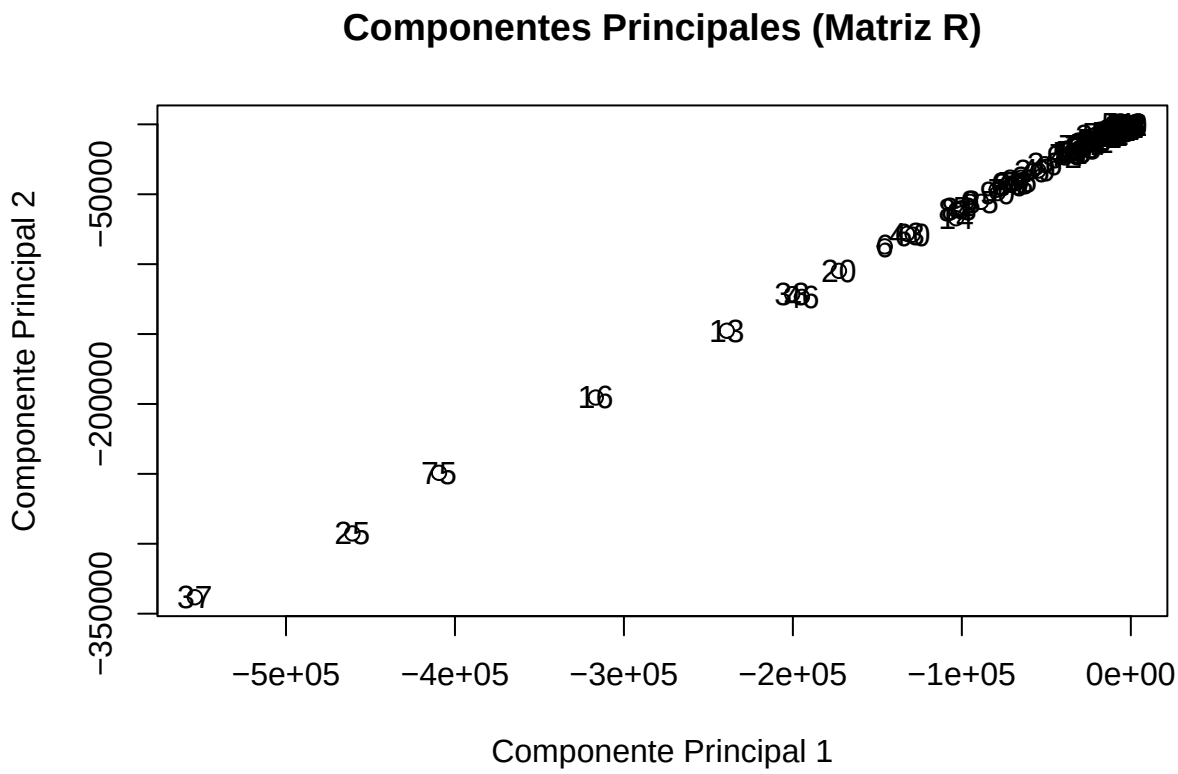


```
# Gráfica biplot (S) para visualizar la relación entre las variables originales y las componentes
biplot(cpS, scale = 0)
```

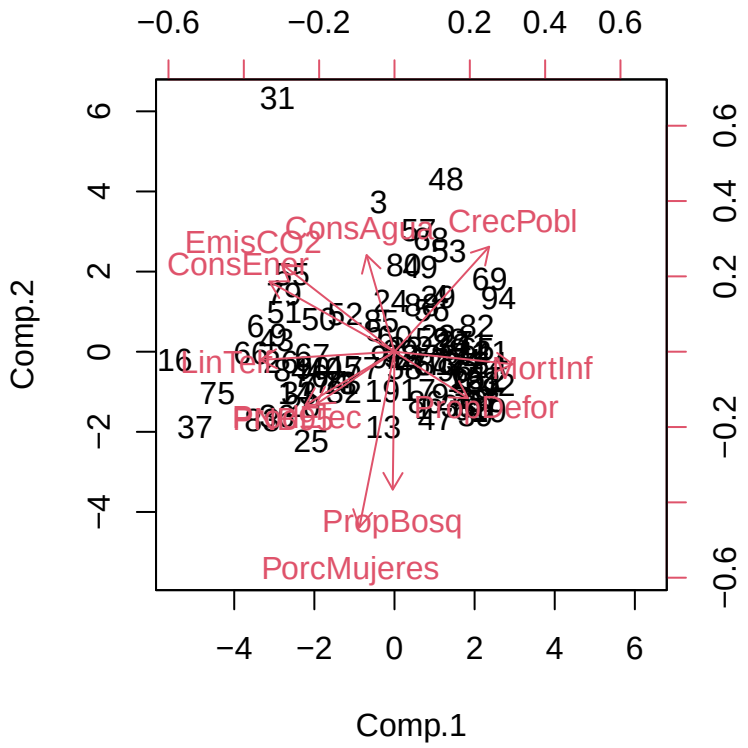
```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```




```r
# Proyectar los datos originales en las dos primeras componentes principales (R)
cpR <- princomp(datos, cor = TRUE)
cpaR <- as.matrix(datos) %*% cpR$loadings

# Gráfica de dispersión de las dos primeras componentes principales (R)
plot(cpaR[, 1], cpaR[, 2], type = "p", main = "Componentes Principales (Matriz R)",
     xlab = "Componente Principal 1", ylab = "Componente Principal 2")
text(cpaR[, 1], cpaR[, 2], labels = row.names(datos))
```




```r
# Gráfica biplot (R) para visualizar la relación entre las variables originales y las componentes
biplot(cpR, scale = 0)
```



Gráfica de Dispersión de las Dos Primeras Componentes Principales (Matriz S):

Esta gráfica muestra cómo se agrupan las observaciones (países) en función de las dos primeras componentes principales. Cada punto representa un país.Los ejes “Componente Principal 1” y “Componente Principal 2” son combinaciones lineales de las variables originales que maximizan la varianza en los datos.Estos resultados sugieren que hay subconjuntos de países que comparten características comunes en nuestras variables económicas y sociales. La “Componente Principal 1” destaca como el principal contribuyente a esta estructura, mientras que la “Componente Principal 2” complementa esta descripción.

Gráfica de Dispersión de las Dos Primeras Componentes Principales (Matriz R):

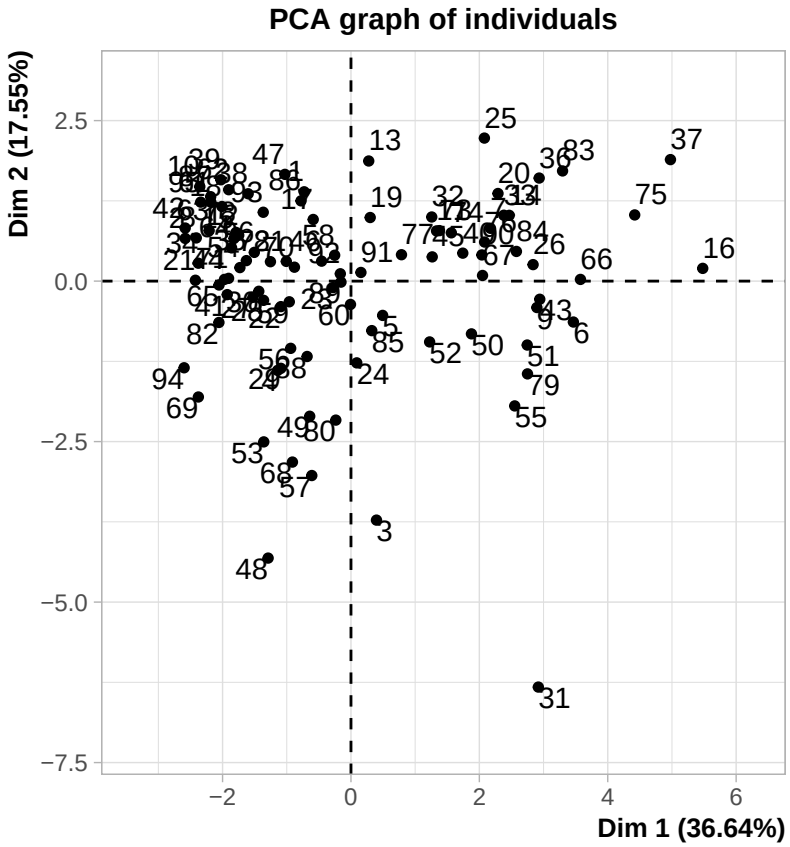
La agrupación de puntos en la esquina superior derecha de la gráfica sugiere que hay un grupo de países que comparten características comunes y se destacan en ambas componentes principales. Esto indica una asociación positiva entre los indicadores económicos y sociales representados por CP1 y CP2.Los resultados del análisis de componentes principales revelan que ciertos indicadores económicos y sociales están correlacionados y muestran un patrón sistemático en los países estudiados. Específicamente, aquellos países que tienen valores más altos en CP1 también tienden a tener valores más altos en CP2, y viceversa.

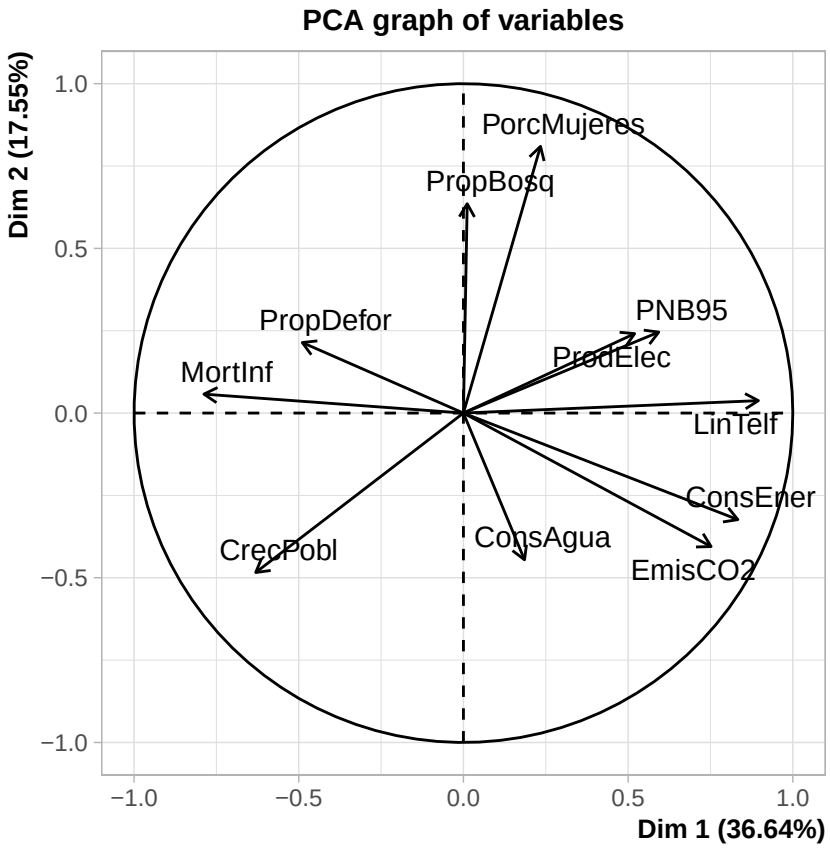
PARTE III

Explore los siguientes gráficos relativos al problema y Componentes Principales y dé una interpretación de cada gráfico.

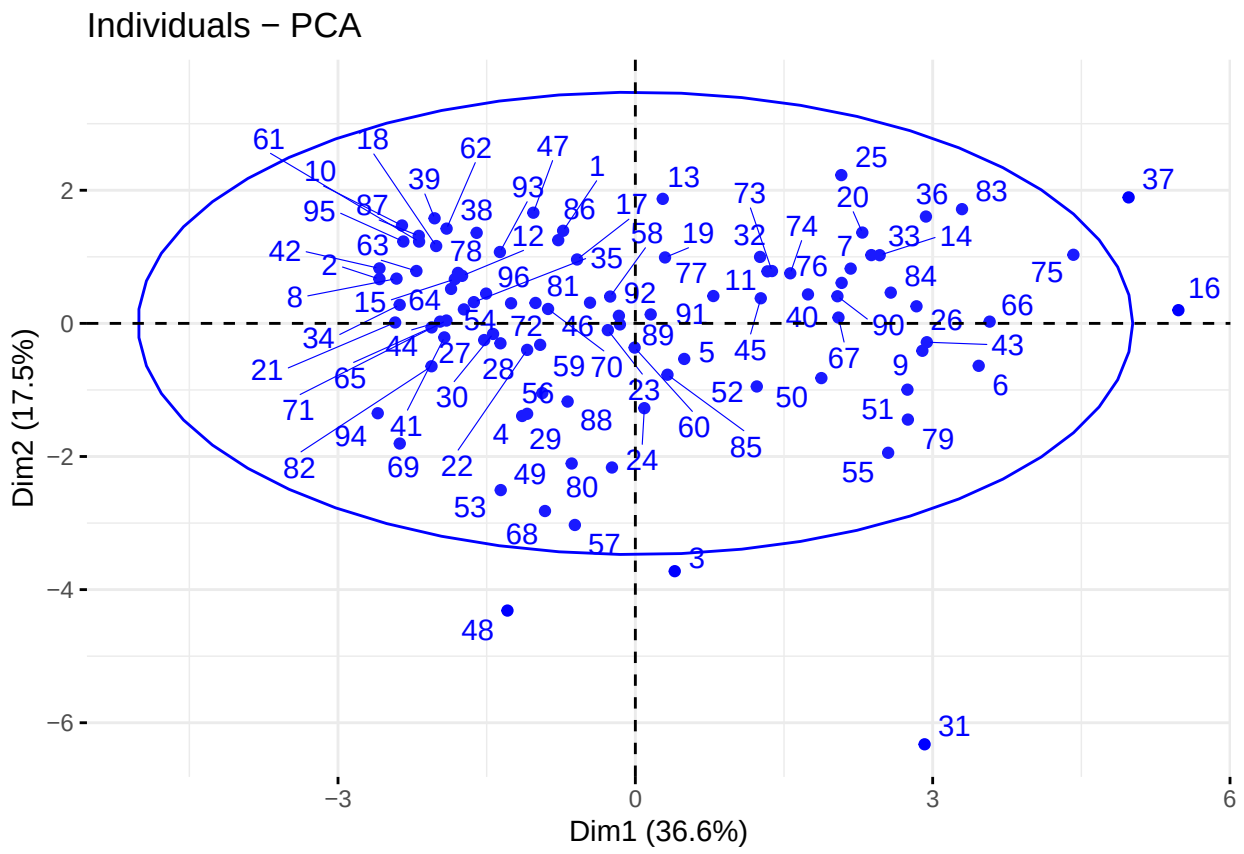
```
library(FactoMineR)
library(factoextra)
library(ggplot2)
datos=M
# Realiza el análisis de Componentes Principales (PCA)
```

```
cp3 <- PCA(datos)
```

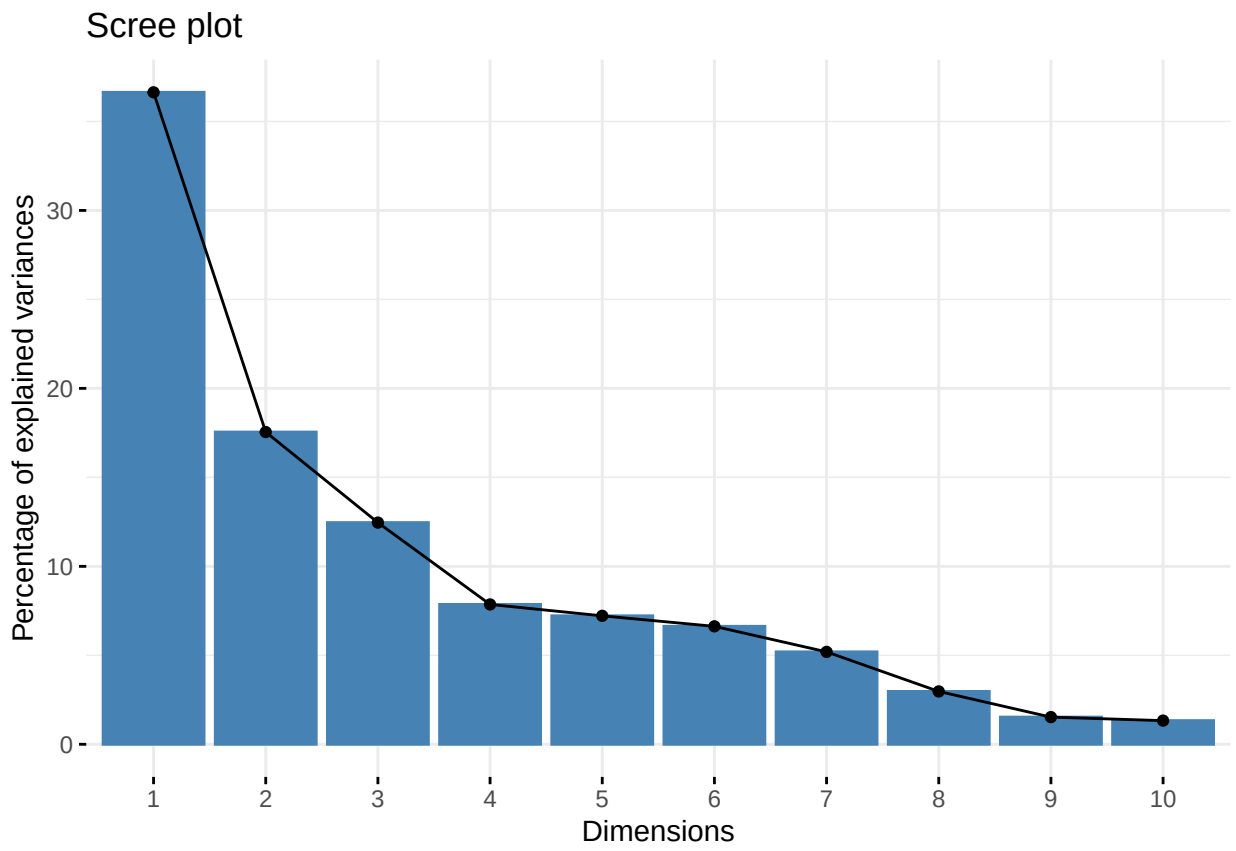




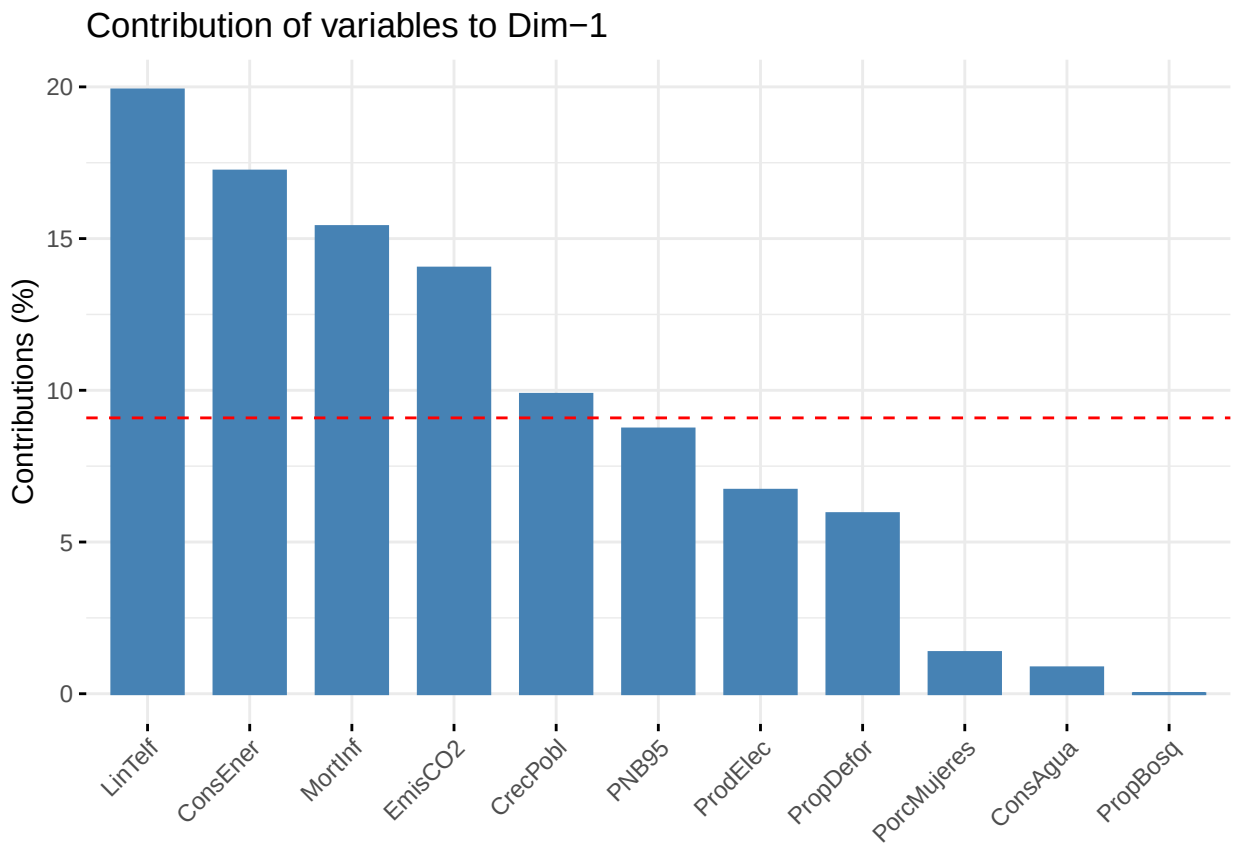

```
# Gráfico de individuos (puntos)
fviz_pca_ind(cp3, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




```
# Gráfico del screeplot para visualizar la varianza explicada
fviz_screeplot(cp3)
```




```
# Gráfico de contribución de variables a las componentes principales
fviz_contrib(cp3, choice = "var")
```



Gráfica de Individuos (PCA Individual Plot):

Esta gráfica muestra cómo se distribuyen los individuos (en este caso, los países) en el espacio definido por las dos primeras componentes principales. Cada punto representa un país y su posición en el gráfico está determinada por los valores que tiene en estas dos componentes principales. Observamos que algunos puntos están más cerca entre sí, formando grupos o clusters. Esto indica que los países que pertenecen a un mismo grupo tienen perfiles socioeconómicos similares en función de las variables analizadas.Algunos puntos están más dispersos, lo que significa que esos países tienen perfiles más diversos en términos de sus indicadores económicos y sociales. La dispersión refleja la variabilidad en los datos.

Gráfica de Scree Plot (Gráfica de Codo):

Esta gráfica muestra la proporción de varianza explicada por cada componente principal. Cada barra representa una componente principal, y su altura indica cuánta varianza explica. En esta gráfica, observamos que la primera componente principal explica la mayor parte de la varianza, mientras que la segunda componente principal también tiene una contribución significativa. A medida que avanzamos en las componentes, la proporción de varianza explicada disminuye, lo que indica que las primeras dos componentes son las más importantes para capturar la variabilidad en nuestros datos.

Gráfica de Contribuciones de Variables:

En esta gráfica, se representan las contribuciones de las variables a las dos primeras componentes principales. Cada variable se muestra como un punto en el gráfico, y su posición refleja cuánto contribuye a estas componentes. Las variables que están más cerca de las componentes principales tienen una contribución significativa a esas componentes. Aquellas que están más alejadas tienen una contribución menor. Esta gráfica nos permite identificar las variables que más influyen en las dos primeras componentes principales y, por lo tanto, tienen un mayor peso en la explicación de la variabilidad observada en los países.