Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad 2. Componentes Principales

TC3007C.501 Inteligencia artificial avanzada para la ciencia de datos II

Profesores:

Iván Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernández

Félix Ricardo Botello Urrutia

Edgar Covantes Osuna

Felipe Castillo Rendón

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

28 de Septiembre de 2023

Actividad 2: Componentes Principales

Parte I

Instrucciones

A partir de los datossobre indicadores económicos y sociales de 96 países "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(stats)
library(factoextra)
```

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

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

```
library(FactoMineR)
library(factoextra)
library(ggplot2)
```

Sugerencias

1. **Calcule las matrices de varianza-covarianza S con cov(X) y la matriz de correlaciones con cor(X)**

```
# Cargar los datos desde el archivo "paises_mundo.csv" (ajusta la ruta según tu ubicación)
datos <- read.csv("paises_mundo.csv")

# Calcular la matriz de varianza-covarianza
matriz_cov <- cov(datos)

cat("Matriz de Covarianza\n", matriz_cov)
```

```
## Matriz de Covarianza
##  1.538298 21.95026 -6.078026 -89333.79 -49739.64 -136.9079 -48.27092 -3.887018 0.3361974 -838.4169 -
```

```
# Calcular la matriz de correlaciones
matriz_corr <- cor(datos)

cat("\nMatriz de Correlación\n", matriz_corr)
```

```
##
## Matriz de Correlación
##  1 0.5506795 -0.5585271 -0.3221215 -0.2971112 -0.5632123 -0.06772953 -0.1565028 0.2010788 -0.3018773
```

2. **Calcule los valores y vectores propios de cada matriz.La función en R es: eigen().**

```
# Calcular los valores y vectores propios de la matriz de varianza-covarianza
eigen_cov <- eigen(matriz_cov)

cat("Valores y Vectores propios de Matriz de Covarianza\n")
```

```
## Valores y Vectores propios de Matriz de Covarianza
```

```
eigen_cov
```

```
## eigen() decomposition
## $values
##  [1] 6.163576e+10 6.581612e+09 4.636256e+06 3.107232e+05 1.216015e+04
```

```
##  [6] 5.137767e+02 3.627885e+02 4.542082e+01 5.800868e+00 1.438020e+00
## [11] 4.768083e-01
##
## $vectors
##                [,1]          [,2]          [,3]          [,4]          [,5]
##  [1,] -1.658168e-06  4.706785e-07  0.0001263736 -1.928408e-05 -0.0055373971
##  [2,] -4.048139e-05 -1.774254e-05  0.0082253821 -2.493257e-03 -0.0944030204
##  [3,]  5.739096e-06 -1.084543e-05  0.0001318149  5.538307e-03  0.0314036410
##  [4,]  8.880376e-01  4.597632e-01  0.0026022071 -3.893588e-04 -0.0003327409
##  [5,]  4.597636e-01 -8.880405e-01  0.0005694896  1.096305e-03  0.0002207819
##  [6,]  3.504341e-04  4.016179e-04 -0.0619424889  7.641174e-03  0.9921404486
##  [7,]  2.625508e-04 -1.122118e-03 -0.0401453227 -9.991411e-01  0.0057795144
##  [8,]  4.089564e-06  7.790843e-06  0.0012719918  6.435797e-03  0.0419331615
##  [9,] -1.073825e-06  2.350808e-07  0.0001916177  4.043796e-05 -0.0018090751
## [10,]  2.547156e-03  7.126782e-04 -0.9972315499  3.973568e-02 -0.0625729475
## [11,]  4.643724e-06 -1.315731e-06 -0.0020679047 -5.626049e-05 -0.0042367120
##                [,6]          [,7]          [,8]          [,9]         [,10]
##  [1,]  1.243456e-02  5.359089e-03 -8.390810e-02 -6.778358e-02 -1.158091e-01
##  [2,]  9.917515e-01  2.258019e-02 -7.891128e-02 -1.637836e-02  4.264872e-04
##  [3,]  8.552991e-02 -1.136481e-01  9.856498e-01 -1.468464e-02  8.241465e-03
##  [4,] -8.621005e-06 -7.566477e-06  1.217248e-05 -3.971469e-07  4.274451e-07
##  [5,]  1.955408e-05  1.544658e-05 -2.558998e-05  1.059471e-06 -1.353881e-06
##  [6,]  9.109622e-02  4.748682e-02 -3.416812e-02 -5.379549e-03 -3.409423e-03
##  [7,] -1.087229e-03 -6.863294e-03  4.698731e-03  7.965261e-05  3.621425e-05
##  [8,]  1.721948e-02 -9.920538e-01 -1.169638e-01  1.416566e-03  5.891758e-03
##  [9,]  1.758667e-03 -7.455427e-03  1.811443e-02  1.283039e-01 -9.859317e-01
## [10,]  2.639673e-03 -3.764707e-03  1.267052e-03  2.262931e-03  2.672618e-04
## [11,] -1.877994e-02 -1.709137e-03 -5.204823e-03 -9.891529e-01 -1.200519e-01
##               [,11]
##  [1,]  9.872887e-01
##  [2,] -2.092491e-02
##  [3,]  8.344324e-02
##  [4,]  2.723996e-07
##  [5,] -2.086857e-07
##  [6,]  4.944397e-04
##  [7,]  4.780416e-04
##  [8,] -3.748976e-03
##  [9,] -1.052934e-01
## [10,]  5.906241e-05
## [11,] -8.221371e-02
```

```r
# Calcular los valores y vectores propios de la matriz de correlaciones
eigen_corr <- eigen(matriz_corr)

cat("\nValores y Vectores propios de Matriz de Correlación\n")
```

```
##
## Valores y Vectores propios de Matriz de Correlación
```

```r
eigen_corr
```

```
## eigen() decomposition
## $values
##  [1] 4.02987902 1.92999195 1.37041115 0.86451597 0.79414057 0.72919997
##  [7] 0.57130511 0.32680096 0.16806846 0.14632819 0.06935866
##
```

```
## $vectors
##               [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
##  [1,] -0.314119414  0.34835747 -0.07352541 -0.44028717 -0.32972147 -0.18392437
##  [2,] -0.392395442 -0.04136238 -0.17759254 -0.13398483  0.08340489 -0.08656390
##  [3,]  0.116546319 -0.58283641  0.16686305  0.05865031  0.18654100  0.16835650
##  [4,]  0.295393771 -0.17690839 -0.53343025 -0.26248209 -0.14110658  0.04653378
##  [5,]  0.258964724 -0.17356372 -0.61438847 -0.17389644 -0.07521971  0.02821905
##  [6,]  0.446082934 -0.02719077  0.15177250  0.04959796 -0.05416498  0.02442175
##  [7,]  0.092410503  0.32060987 -0.37024258  0.73603097  0.02671021 -0.30940890
##  [8,]  0.005692925 -0.45742697  0.16480339  0.04024882 -0.41531702 -0.75356463
##  [9,] -0.243652293 -0.15408201 -0.02961449  0.33650345 -0.73261463  0.50894232
## [10,]  0.415029554  0.23286257  0.20608749 -0.06730166 -0.23100421  0.05806466
## [11,]  0.374531032  0.29168698  0.20631751 -0.14843513 -0.24028756 -0.02809233
##               [,7]        [,8]        [,9]       [,10]       [,11]
##  [1,]  0.1628974320 -0.09481963 -0.52181220  0.34674573 -0.10062784
##  [2,]  0.6398040762 -0.32307802  0.29031618 -0.38959240  0.17487096
##  [3,]  0.5310867107  0.05209889 -0.23599758  0.42854658 -0.16786800
##  [4,] -0.1490207046 -0.44913216  0.36995675  0.34911534 -0.15247432
##  [5,]  0.1082745817  0.50343911 -0.30681318 -0.33770404  0.12366382
##  [6,] -0.0008501608 -0.56975094 -0.44733110 -0.20997673  0.44992596
##  [7,]  0.2357666690 -0.05962470 -0.08358225  0.20561803 -0.07067780
##  [8,] -0.0806036686  0.04275404  0.07438520 -0.08671232 -0.01493710
##  [9,]  0.0112333588 -0.01607505  0.01868615 -0.03209758  0.07259619
## [10,]  0.2711228006 -0.05023582  0.04339752 -0.36147417 -0.67912543
## [11,]  0.3352822144  0.30978009  0.37666244  0.28779437  0.46737561
```

3. **Calcule la proporción de varianza explicada por cada componente. Se sugiere dividir cada lambda entre la varianza total (las lambdas están en eigen(S)[1]). La varianza total es la suma de las varianzas de la diagonal de S. Una forma es sum(diag(S)). La varianza total de los componentes es la suma de los valores propios (es decir, la suma de la varianza de cada componente), sin embargo, si sumas la diagonal de S (es decir, la varianza de cada x), te da el mismo valor (¡comprúebalo!). Recuerda que las combinaciones lineales buscan reproducir la varianza de X.**

```r
# Calcular la proporción de varianza explicada para la matriz de varianza-covarianza
prop_var_explicada_cov <- eigen_cov$values / sum(eigen_cov$values)

cat("Varianza explicada para Matriz de Covarianza\n", prop_var_explicada_cov)
```

```
## Varianza explicada para Matriz de Covarianza
##  0.9034543 0.09647298 6.795804e-05 4.554567e-06 1.782429e-07 7.530917e-09 5.317738e-09 6.657763e-10 8
```

```r
# Calcular la proporción de varianza explicada para la matriz de correlaciones
prop_var_explicada_corr <- eigen_corr$values / sum(eigen_corr$values)

cat("\nVarianza explicada para Matriz de Correlación\n", prop_var_explicada_corr)
```

```
##
## Varianza explicada para Matriz de Correlación
##  0.3663526 0.1754538 0.1245828 0.07859236 0.0721946 0.06629091 0.05193683 0.02970918 0.01527895 0.013
```

4. **Acumule los resultados anteriores. (cumsum() puede servirle).**

```r
# Acumular las proporciones de varianza explicada
var_acumulada_cov <- cumsum(prop_var_explicada_cov)

print("Acumulación de proporciones de Varianza Explicada para Matriz de Covarianza\n")
```

```
## [1] "Acumulación de proporciones de Varianza Explicada para Matriz de Covarianza\n"
```

```
var_acumulada_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 1.0000000
```

```
var_acumulada_corr <- cumsum(prop_var_explicada_corr)

print("\nAcumulación de proporciones de Varianza Explicada para Matriz de Correlación\n")
```

```
## [1] "\nAcumulación de proporciones de Varianza Explicada para Matriz de Correlación\n"
```

```
var_acumulada_corr
```

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

5. **Según los resultados anteriores, ¿qué componentes son los más importantes? ¿qué variables son las que más contribuyen a la primera y segunda componentes principales? ¿por qué lo dice? ¿influyen las unidades de las variables?**

6. **Hacer los mismos pasos anteriores, pero con la matriz de correlaciones (se obtiene con cor(x) si x está compuesto por variables numéricas)**

En el análisis de la matriz de covarianzas, la primera componente principal es la más importante, ya que explica aproximadamente el 90.34% de la varianza total de los datos. En el análisis de la matriz de correlaciones, la primera componente principal también es la más importante, explicando aproximadamente el 36.64% de la varianza total.

Para la primera componente principal en ambos análisis (matriz de covarianzas y matriz de correlaciones), las variables que más contribuyen son aquellas con los mayores valores absolutos en los vectores propios correspondientes a esta componente. En el caso de la matriz de covarianzas, estas variables incluyen las columnas 4, 5 y 7, que son las más influyentes en la primera componente. En la matriz de correlaciones, las variables más influyentes en la primera componente son las columnas 1, 2, y 3.

Para la segunda componente principal, se aplicaría el mismo principio. Sin embargo, en ambos análisis, la segunda componente tiene una varianza explicada mucho menor que la primera, lo que indica que su importancia es menor.

Las unidades de las variables influyen en los resultados de los componentes principales, especialmente si las variables tienen escalas muy diferentes. En este caso, al observar los vectores propios (loadings), es probable que las variables con unidades más grandes tengan loadings más grandes debido a su mayor varianza. Esto puede hacer que las variables con unidades más grandes parezcan más influyentes en las componentes principales. Por lo tanto, la estandarización de las variables (restar la media y dividir por la desviación estándar) puede ser importante para tratar todas las variables en la misma escala antes de realizar el PCA. Si no se estandarizaron las variables, la interpretación de la importancia de las variables debe considerar el impacto de las unidades.

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

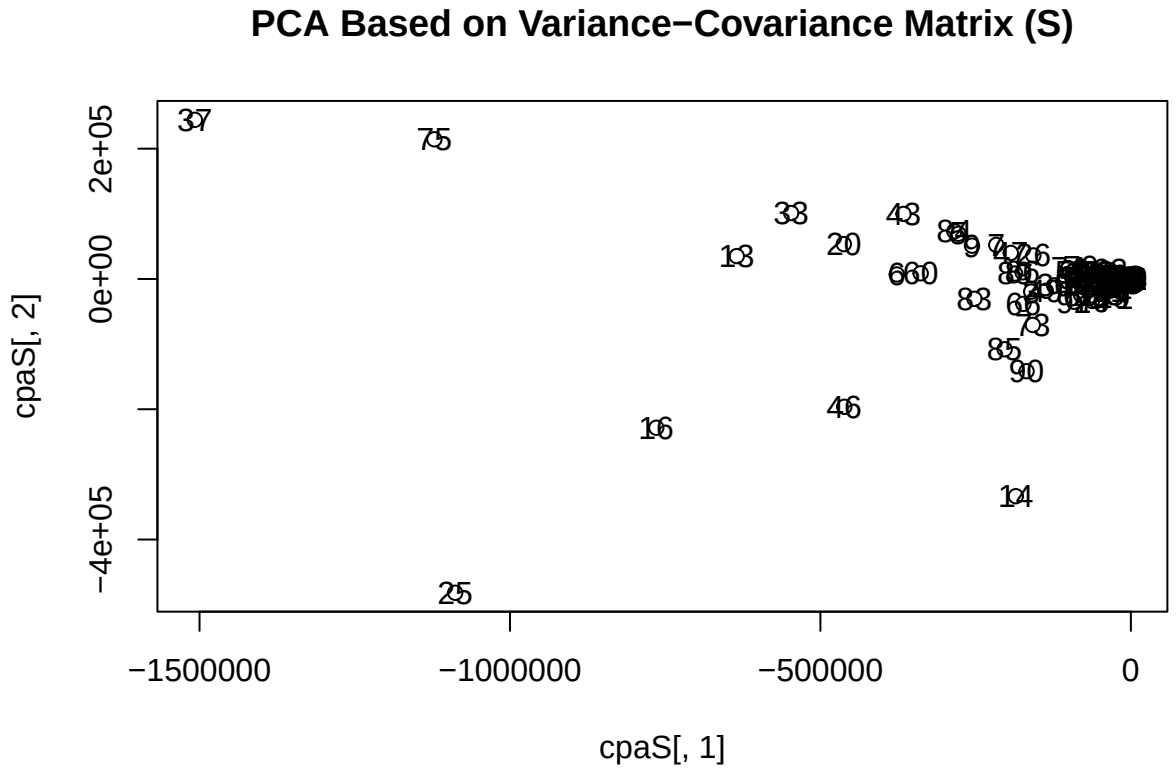
La matriz de varianza-covarianza explica más de la varianza en los datos con los primeros componentes en comparación con la matriz de correlación. La varianza acumulada explicada por el primer componente es significativamente mayor en la matriz de varianza-covarianza. Esto sugiere que las variables originales están altamente correlacionadas y que el uso de la matriz de correlación podría no proporcionar tanta reducción de dimensionalidad.

Parte II

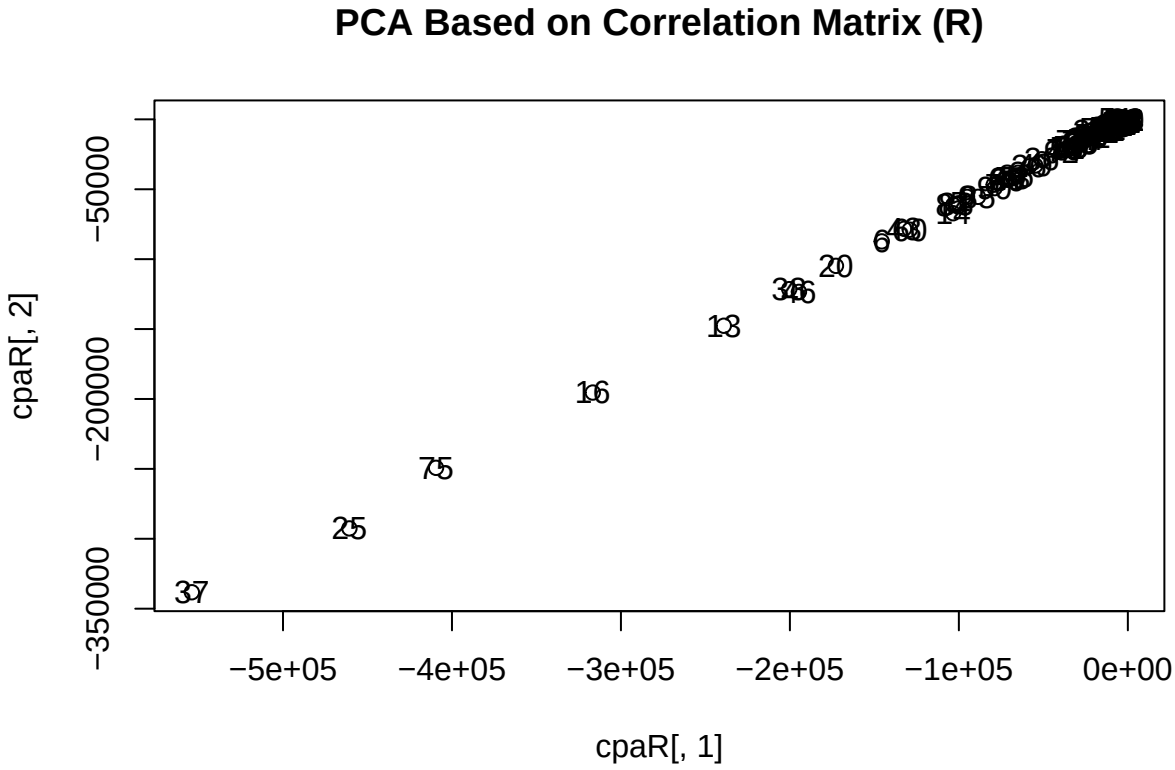
Instrucciones

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")

```
# Utilizamos la función PCA para las matrices S y R
cpS <- princomp(datos, cor = FALSE)  # PCA utilizando la matriz varianza-covarianza (S)
cpR <- princomp(datos, cor = TRUE)   # PCA utilizando la matriz de correlación (R)

# Proyectar datos hacia los primeros 2 componentes para ambas matrices
cpaS <- as.matrix(datos) %*% cpS$loadings
cpaR <- as.matrix(datos) %*% cpR$loadings

# Creamos un diagrama de dispersión para los primeros 2 componentes de la matriz S
plot(cpaS[, 1], cpaS[, 2], type = "p", main = "PCA Based on Variance-Covariance Matrix (S)")
text(cpaS[, 1], cpaS[, 2], 1:nrow(cpaS))
```

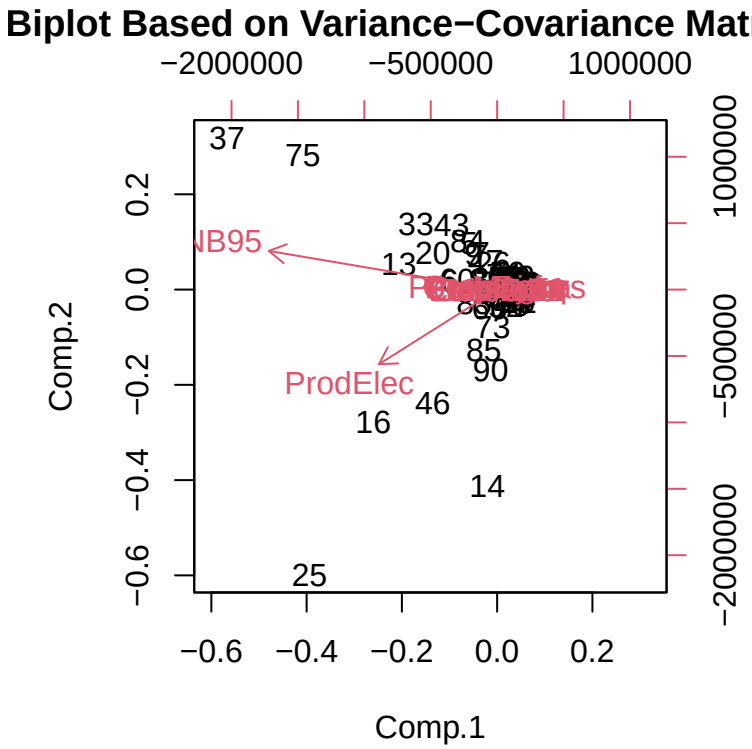



```
# Creamos un diagrama de dispersión para los primeros 2 componentes de la matriz R
plot(cpaR[, 1], cpaR[, 2], type = "p", main = "PCA Based on Correlation Matrix (R)")
text(cpaR[, 1], cpaR[, 2], 1:nrow(cpaR))
```

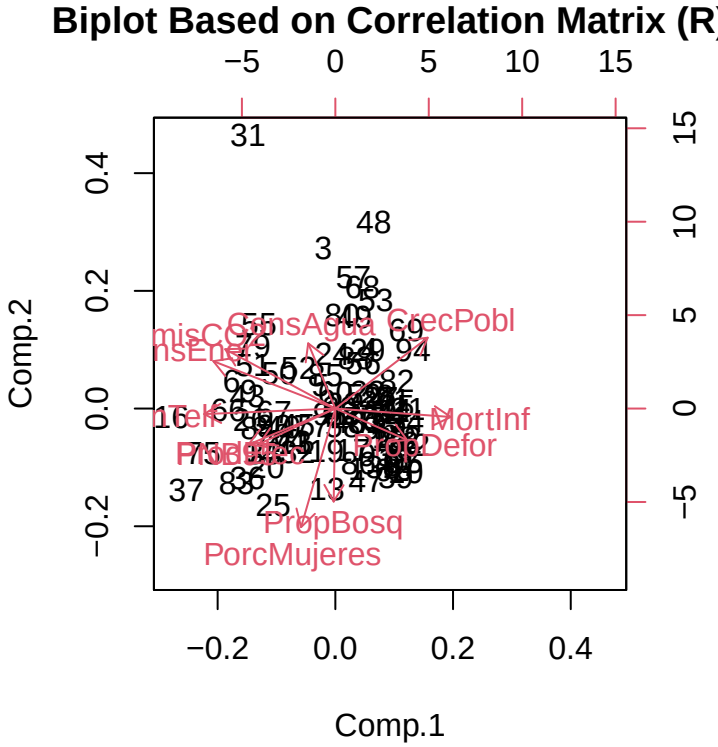


```r
# Creamos una biplot para ambas matrices
biplot(cpS, main = "Biplot Based on Variance-Covariance Matrix (S)")
```

```
## 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
```




```
biplot(cpR, main = "Biplot Based on Correlation Matrix (R)")
```



Parece que el análisis de la matriz de covarianza muestra un factor dominante (Comp. 1) que explica la mayor parte de la variabilidad en los datos, mientras que el análisis de la matriz de correlación revela un patrón más complejo con múltiples variables que contribuyen a los dos primeros componentes. Esta diferencia entre las dos matrices puede deberse a la escala (varianza contra correlación) y puede proporcionar información sobre cómo se relacionan las variables entre sí.

Algunos puntos a considerar son los siguientes:

- Con respecto a la gráfica de dispersión de la matriz de covarianza, la concentración de datos cerca de (0, 0e+00) sugiere que el primer componente principal (Comp. 1) está capturando una cantidad sustancial de varianza en los datos. Esto podría indicar que un factor dominante está impulsando la variabilidad entre las variables.
- Hablando de la gráfica de dispersión de la matriz de correlación, el comportamiento inclinado de los datos sugiere que existe una relación o tendencia lineal en los datos. A medida que se avanza a lo largo de la pendiente, las variables tienden a moverse juntas de manera coordinada. La concentración de datos a medida que aumenta la pendiente implica que las variables que están correlacionadas positivamente están contribuyendo a esta tendencia.
- Si observamos el biplot de la matriz de covarianza, nos podremos dar cuenta que la concentración de los datos en un punto específico indica que los 2 componentes principales explican una larga porción de la variabilidad en los datos, y la mayoría de las variables se encuentran muy agrupadas alrededor de este mismo punto.

La ligera dispersión de datos que se muestra en este mismo gráfico indica poca variabilidad no explicada por estos 2 componentes. Por último, existen 2 variables (*NB95* y *ProdElec*) que tienen una influencia significativa en ambos componentes. Estos contribuyen a la mayor parte de los componentes principales y están correlacionados de forma negativa entre ellos.

- Finalmente, el biplot de la matriz de correlación presenta una concentración de datos en una misma área, esto indica que los 2 principales componentes basados en la matrix de correlación también explican una porción larga de la variabilidad de los datos. La dispersión existente de algunos de los datos sugieren algo de variabilidad residual no capturada por los 2 componentes.

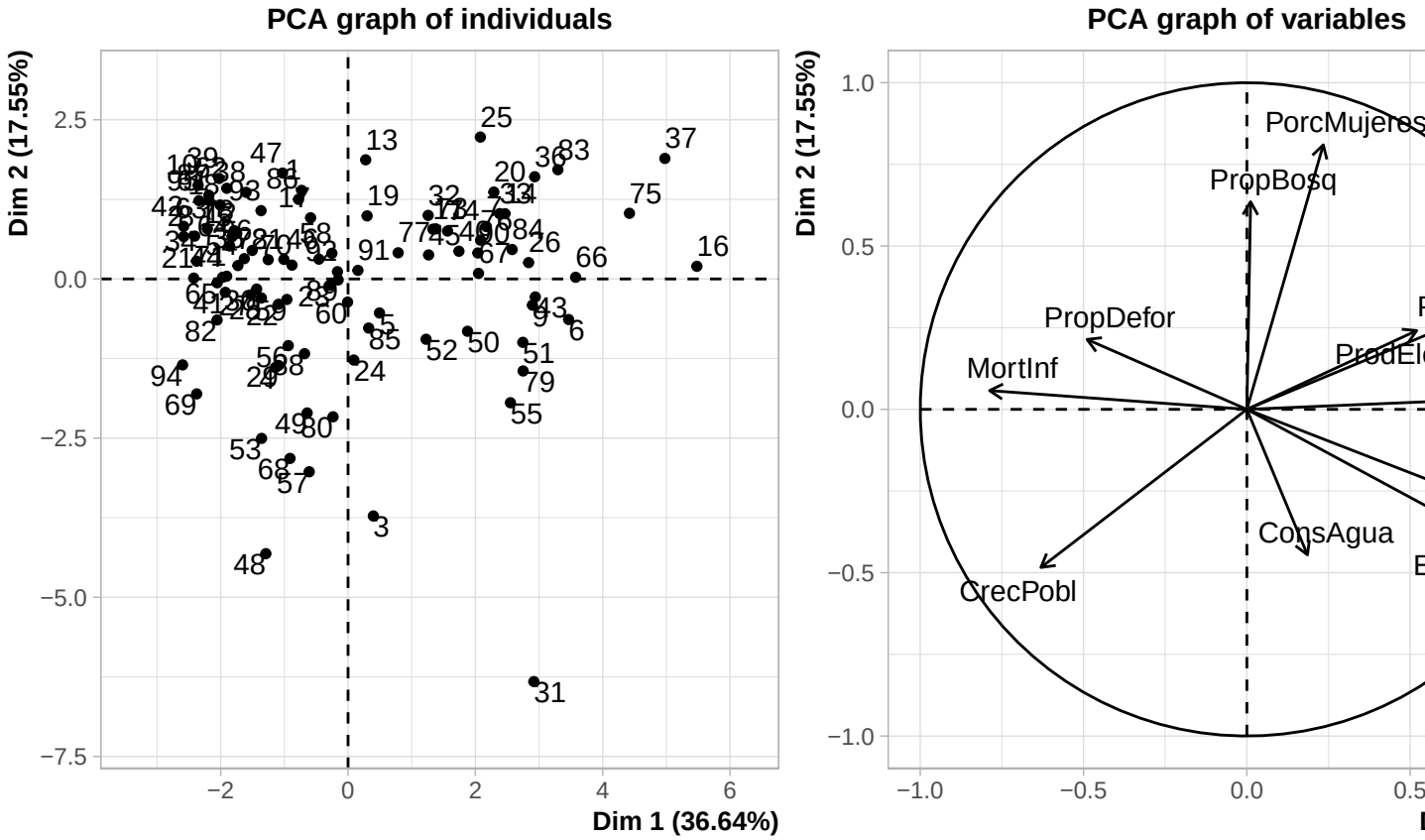
En cuanto a las variables, tenemos bastantes que contribuyen a los componentes principales. Dichas variables poseen relaciones muy complejas, ya que no se encuentran apuntando a las mismas direcciones, cada una tiene su dirección y magnitud propias, distinto del resto, esto también demostrando tanto correlaciones positivas como negativas.

Parte III

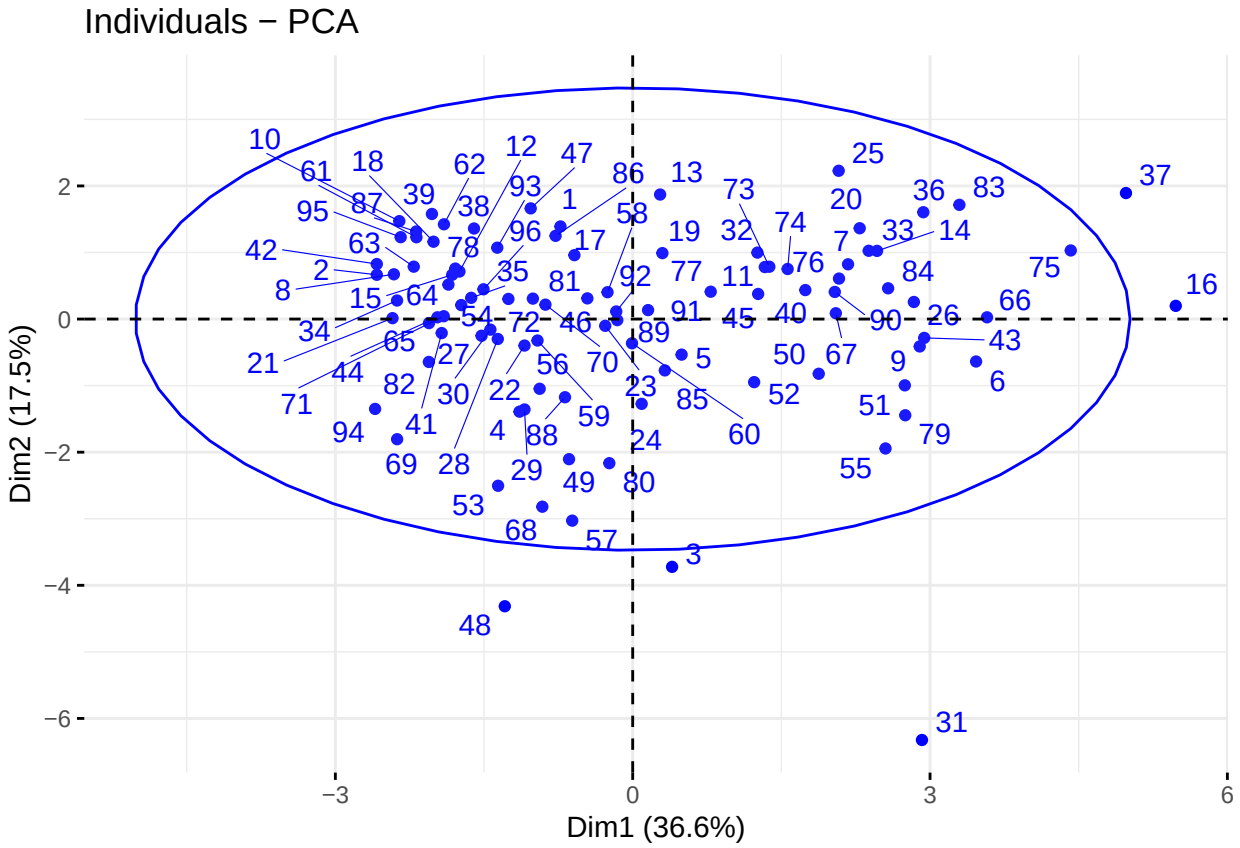
Instrucciones

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

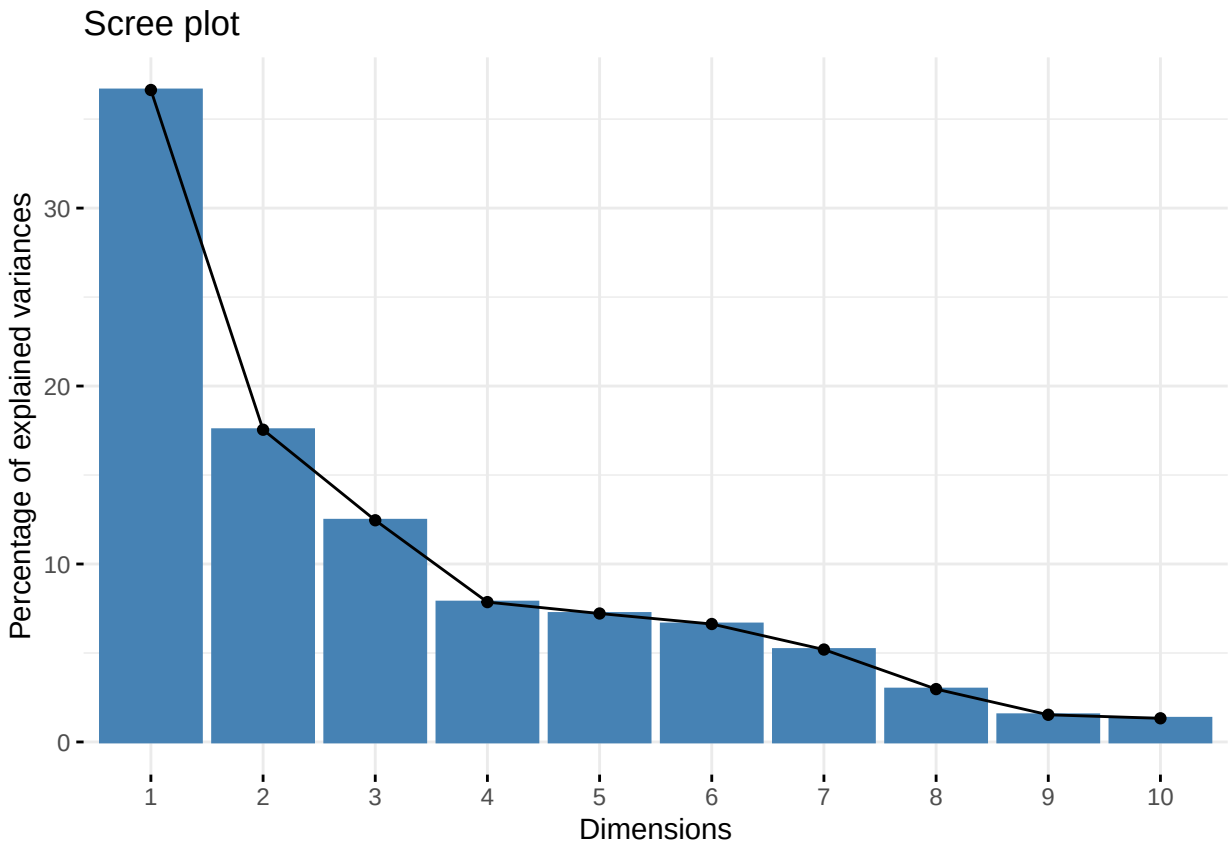
```
cp3 = PCA(datos)
```



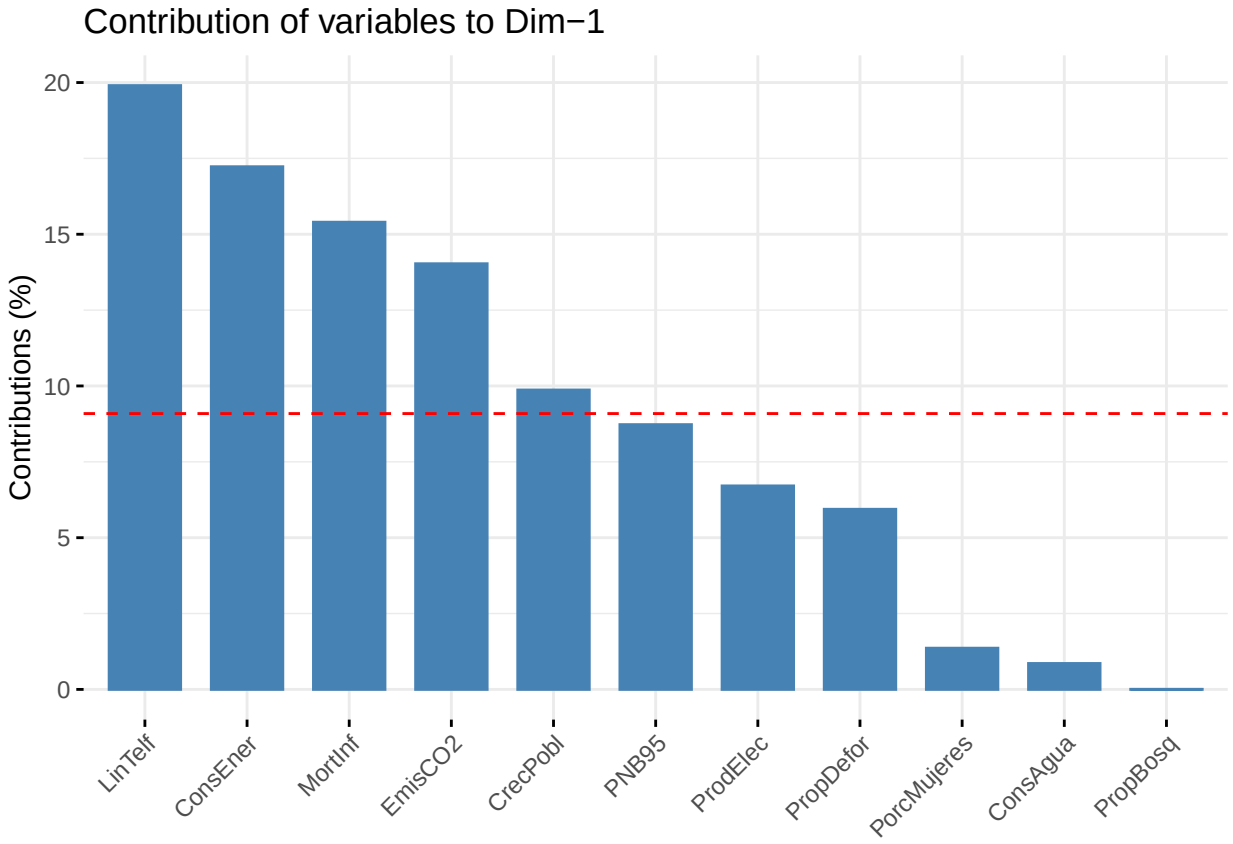

```
fviz_pca_ind(cp3, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




```
fviz_screeplot(cp3)
```




```
fviz_contrib(cp3, choice = c("var"))
```



1. Como podemos observar en la primera gráfica, los puntos que representan cada uno a los países, entre más cerca se encuentran uno del otro, más características en común presentan, ya sean indicadores sociales o económicos.
2. Este gráfico es muy parecido al visto anteriormente en la parte anterior de esta actividad, y es que en base a la magnitud y dirección de las flechas que representan cada variable del dataset, nos permite ver cuál de estas aporta más a los componentes principales en el contexto de nuestro problema. En base a las direcciones de las flechas, podemos discernir en qué variables se encuentran asociadas positiva o negativamente con cada componente principal.
3. En el tercer gráfico de PCA individuales, los puntos azules representan cada país dentro del dataset y se encuentran proyectados en los 2 componentes principales. Los elipses dibujados alrededor de cada agrupamiento de puntos ayuda a identificar fácilmente grupos o patrones dentro de los datos. Esto quiere decir que los países que se acerquen más entre ellos, poseen mayores características en común, ya sean patrones económicos o indicadores sociales.
4. En cuanto al ScreePlot, este es uan representación gráfica de lso valores propios de los componentes principales. Muestran cuánta varianza explcia cada componente principal. Cuando los valores propios empiezan a decrecer, lo cuál se ve desde la segunda dimension, este punto en específico indica el número de componentes principales que explican unaporción substancial de la variabilidad de los datos, osea, 2 componentes deben preservarse para el análisis.
5. La gráfica de contribución de variables, como su nombre indica, muestra la contribución de cada una de las variables originales con respecto a los componentes principales. En el gráfico de barras, podemos notar que de izquierda a derecha, tenemos las variables que más aportan a las que menos con respecto a los componentes principales. La que más aporta es *LinTelf*, seguido de *ConsEner* y *MortInf*. Y la que menos aporta es *PropBosq*, precedida por *ConsAgua* y *PorcMujeres*