Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

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

A2-Componentes-Principales

Rodolfo Sandoval Schipper A01720253

29 sept 2023

A2 Componentes Principales

Rodolfo Sandoval A01720253

2023-09-30

Parte 1

#Primero Cargamos los datos del archivo CSV

```
data <- read.csv("paises_mundo.csv", header = TRUE)
```

#1. Calculamos la la matriz de covarianza y correlaciones

```
data_numerico <- data[, sapply(data, is.numeric)]
cov_matriz <- cov(data_numerico)
cor_matriz <- cor(data_numerico)
```

#2. Calculamos los vectores para cada matriz

```
eigen_cov <- eigen(cov_matriz)
eigen_cor <- eigen(cor_matriz)
```

#3. Ahora claculamos la proporcion de la varianza para cada componente

```
proporcion_cor <- eigen_cor$values / sum(diag(cor_matriz))
proporcion_cov <- eigen_cov$values / sum(diag(cov_matriz))
print(proporcion_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(proporcion_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
```

#4. Acumulamos los resultados

```
acc_var_cov <- cumsum(proporcion_cov)
acc_var_cor <- cumsum(proporcion_cor)
print(acc_var_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(acc_var_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
```

#5. Determinamos los componentes mas importantes De acuerdo a las variables que se crearon como combinaciones de las variables originales podemos interpretar que en este resumen o analisis en la matriz de covarianza son 2 componentes principales debido a su puntuaje en los resultados. Estas dos variables capturan tendencias en los datos originales y son las dimensiones mas importantes en el resultado. Para la matriz de correlacion, obtenemos una falta de componenetes con alta relacion a las variables, esto es debido a que la matriz de correlacion normaliza las variables. Finalmente, podemos concluir que hay dos componentes importantes que se obtienen de la matriz de covarianza.

#6 y 7 Variables mas influyentes en los primeros componentes principales para ambas matrices

```
cargado_cov <- eigen_cov$vectors[, 1]
importantes_vars_cov <- colnames(data)[order(-abs(cargado_cov))]

cargado_cor <- eigen_cor$vectors[, 1]
importantes_vars_cor <- colnames(data)[order(-abs(cargado_cor))]

print(importantes_vars_cov)
```

```
##  [1] "PNB95"       "ProdElec"    "ConsEner"    "LinTelf"     "ConsAgua"
##  [6] "MortInf"     "PorcMujeres" "EmisCO2"     "PropBosq"    "CrecPobl"
## [11] "PropDefor"
```

```
print(importantes_vars_cor)
```

```
##  [1] "LinTelf"     "ConsEner"    "MortInf"     "EmisCO2"     "CrecPobl"
##  [6] "PNB95"       "ProdElec"    "PropDefor"   "PorcMujeres" "ConsAgua"
## [11] "PropBosq"
```

8. Comparar los resultados

Puedes comparar las proporciones acumulativas y las variables más influyentes entre covarianza y correlación

Se puede observar que en la matriz de covarianza, las variables mas importantes en la primera componente principal son, en orden descendente de importancia -> PNB95, ProdElec, ConsEner, LinTelf. Y en la matriz de correlaciones, las variables mas importantes en la primera componente principal son -> LinTelf, ConsEner, MortInf, EmisCO2. Sin embargo como se menciono anteriormente en el inciso 5 la matriz de covarianza captura tendencias mayores a la matriz de correlacion en los datos originales el cual indica que son las dimensiones mas importantes en el resultado.

##Parte 2

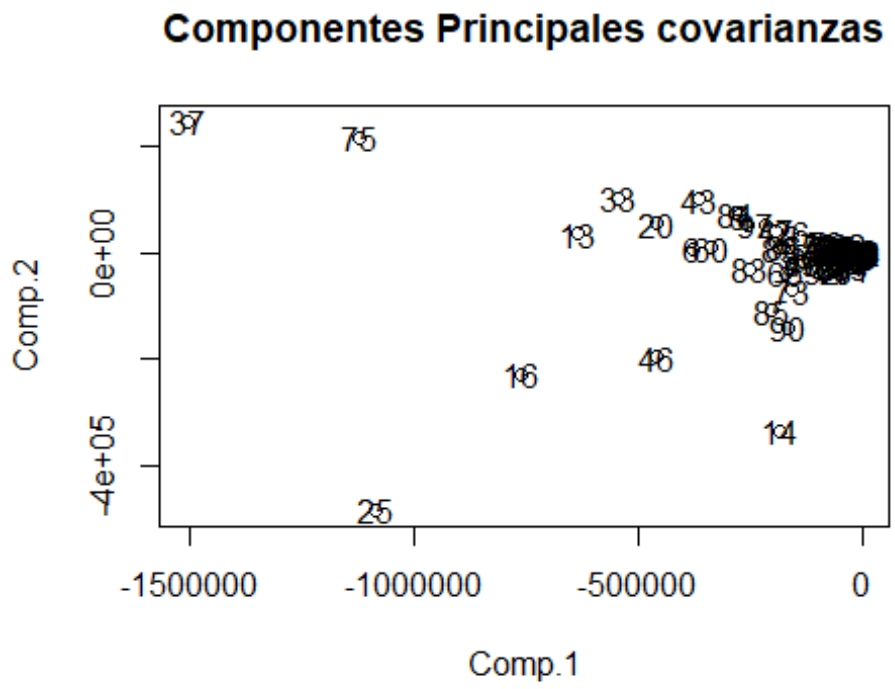
```r
# Cargamos librerias
library(stats)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

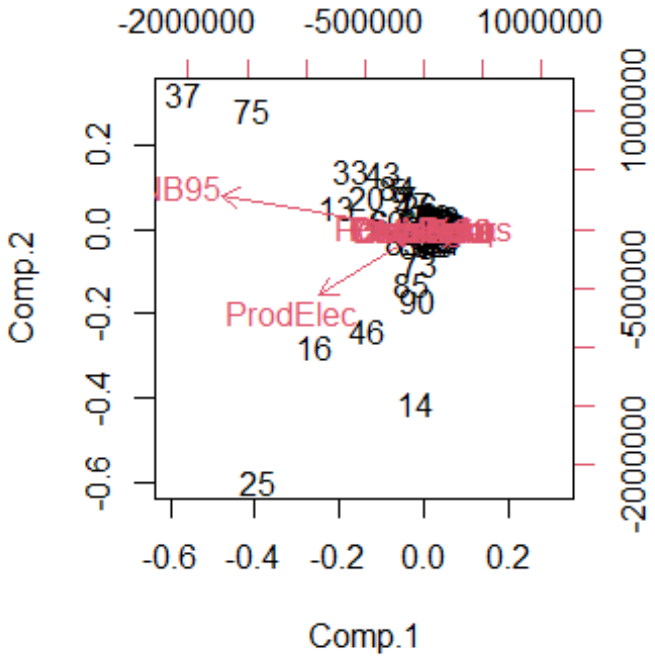
```
## Loading required package: ggplot2
```

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

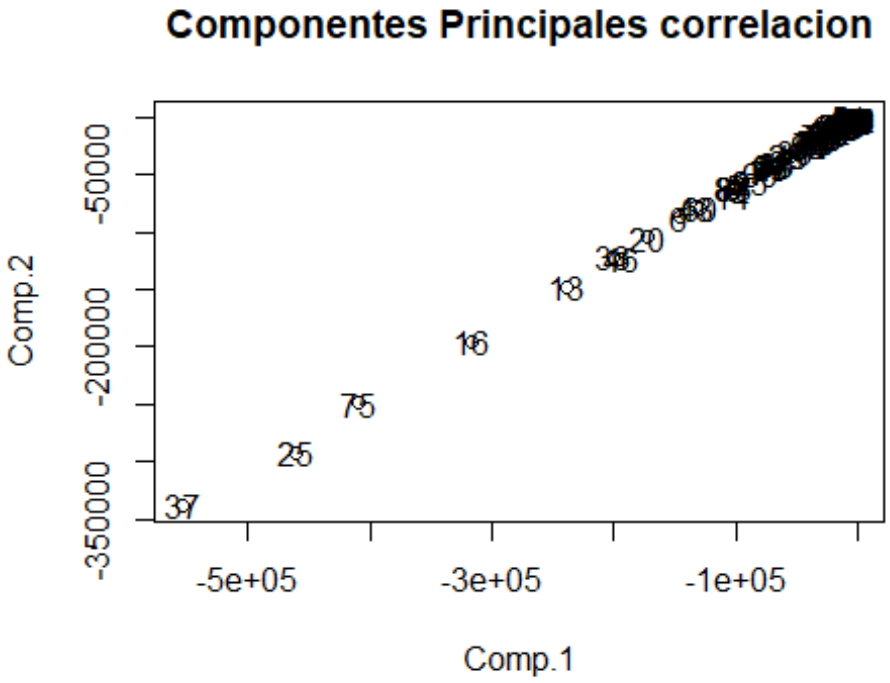
```r
library(ggplot2)

datos <- read.csv("paises_mundo.csv")

# Calcular las componentes principales con la matriz de covarianzas
cpS <- princomp(datos, cor = FALSE)
cpaS <- as.matrix(datos) %*% cpS$loadings


plot(cpaS[, 1:2], type = "p", main = "Componentes Principales covarianzas")
text(cpaS[, 1], cpaS[, 2], labels = 1:nrow(cpaS))
```



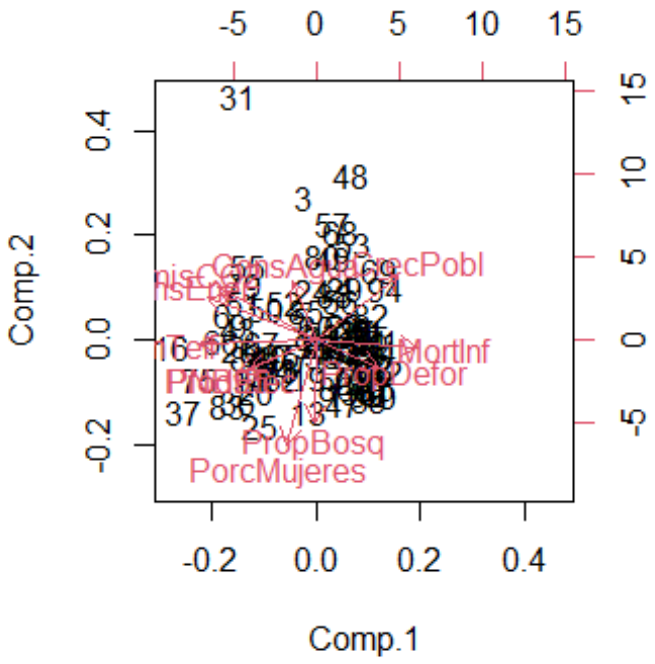

```r
biplot(cpS)
```




```r
# Calcular las componentes principales con la matriz de correlaciones
cpR <- princomp(datos, cor = TRUE)
cpaR <- as.matrix(datos) %*% cpR$loadings

plot(cpaR[, 1:2], type = "p", main = "Componentes Principales correlacion")
text(cpaR[, 1], cpaR[, 2], labels = 1:nrow(cpaR))
```

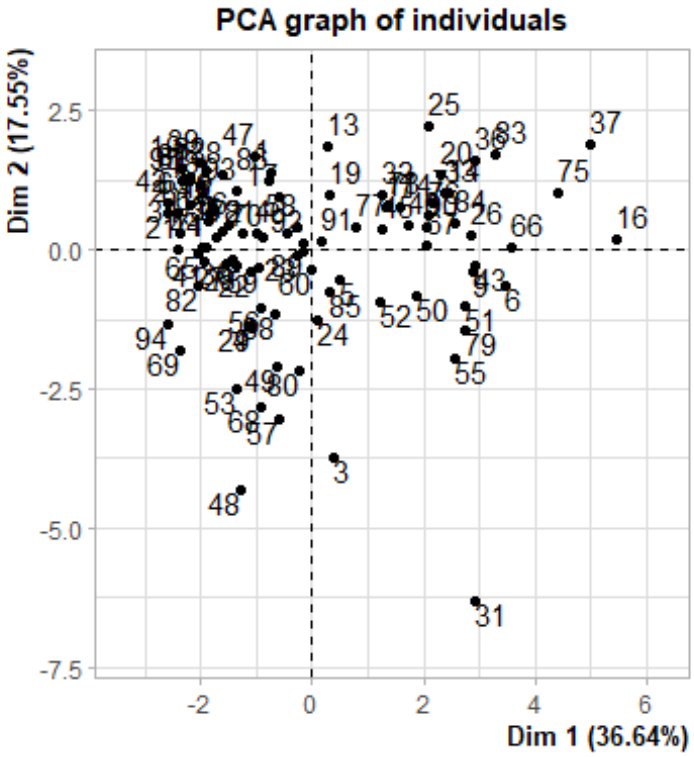

Componentes Principales correlacion
Comp.2
-50000
-200000
-350000
-5e+05
-3e+05
-1e+05
Comp.1

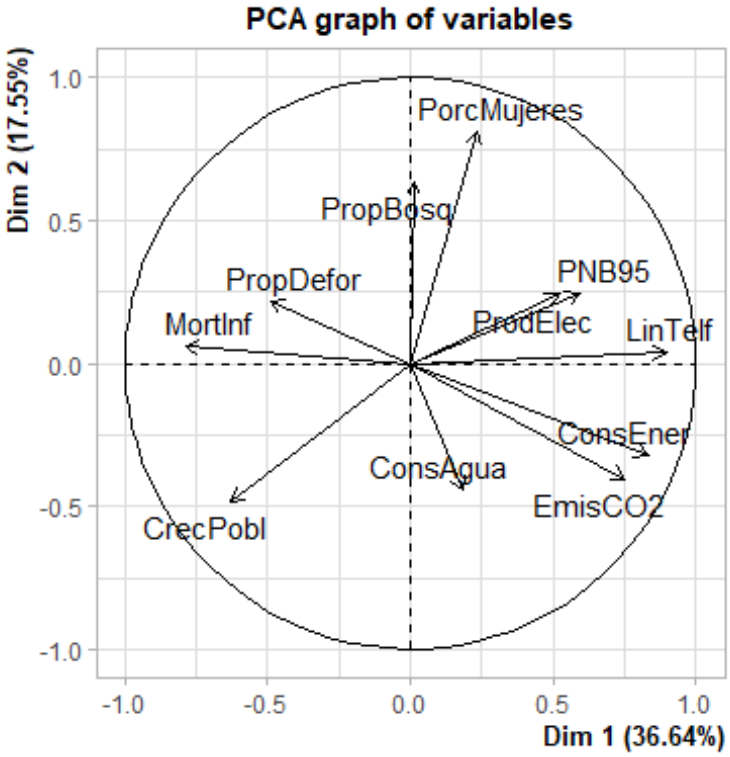
```r
biplot(cpR)
```


-5
0
5
10
15
Comp.2
0.4
0.2
0.0
-0.2
PropBosq
PorcMujeres
-0.2
0.0
0.2
0.4
Comp.1

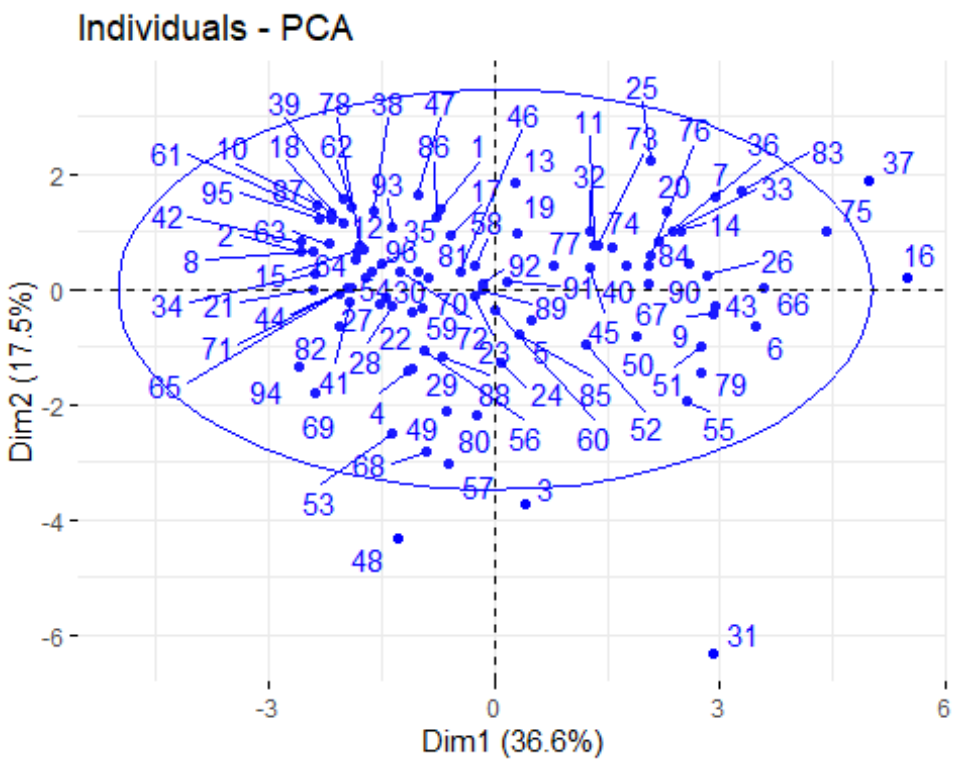
##Parte3

```
# Cargar las bibliotecas
library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.1.3

library(factoextra)
library(ggplot2)

cp3 <- PCA(data)
```

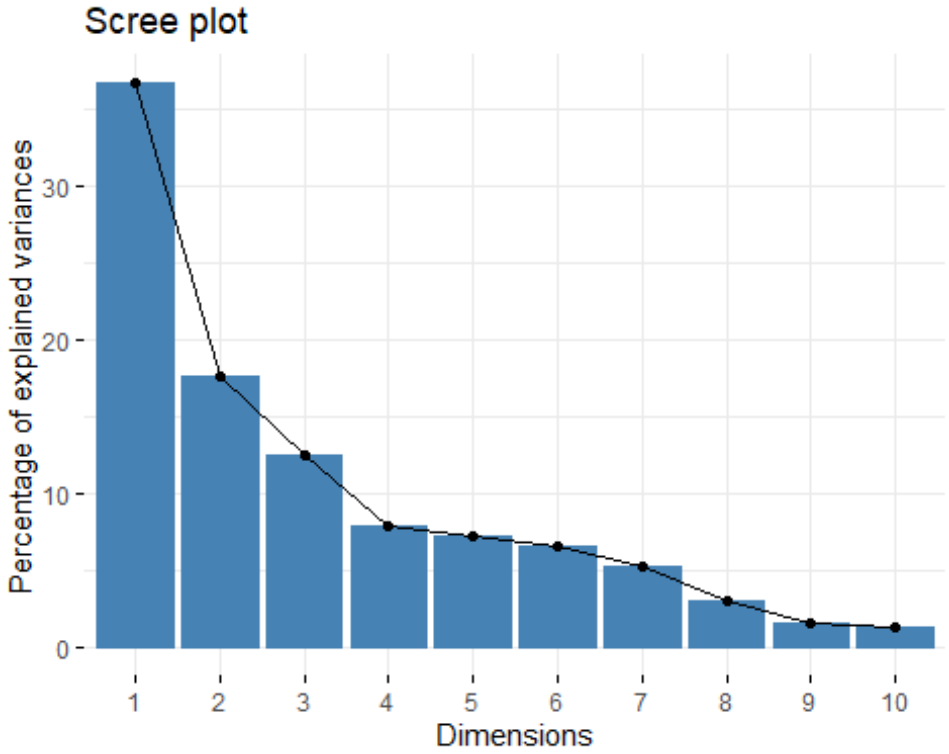




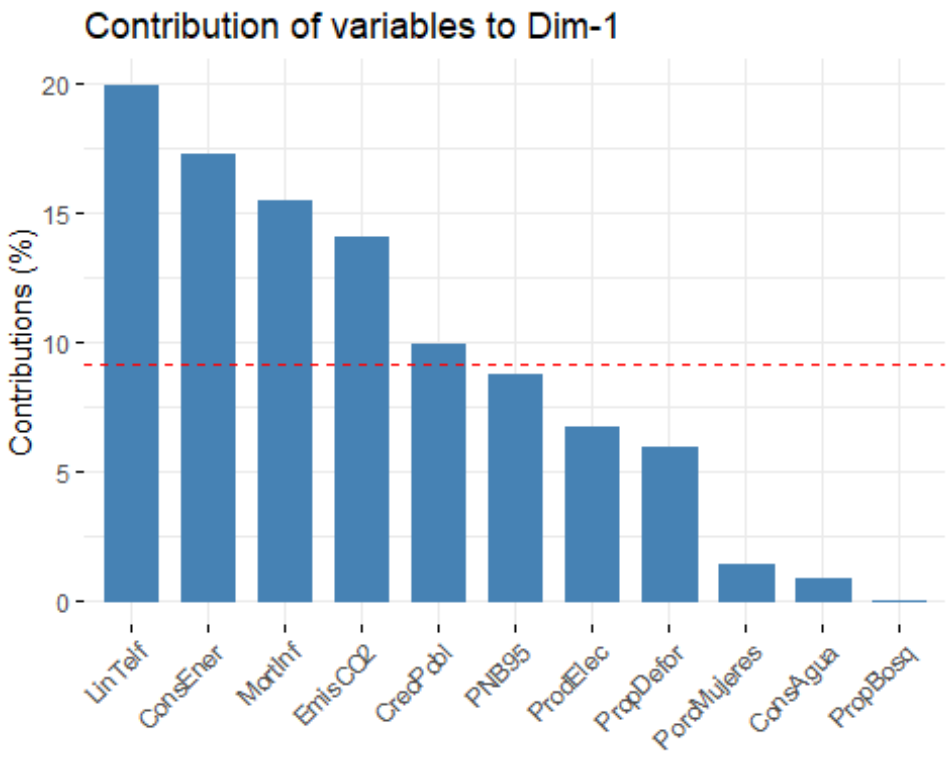

```
fviz_pca_ind(cp3, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




```
fviz_screeplot(cp3)
```




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



#Conclusion e interpretaciones Los resultados a partir del analisis de PCA proporcionan correctamente la informacion sobre la agrupacion de las variables en funcion de su contribucion a los componentes principales. Sin embargo, se analizaron las graficas para entender los resutlados a profundo. El cual se obtuvo que -> El componente principal Comp1 muestra una fuerte contribución positiva de la variable comp3. Y negativas con el resto. La variable Comp9 tiene la contribución positiva más significativa en el componente comp2. Y negativas con 8 y 10. Comp 6 y 10 tienen contribuciones positivas significativas en el componente 3. Y negativas con 1 y 9. En el componenete 4 comp8 tiene una contribucion significativa positiva. Y negativas con 1, 2, y 6. El componente 5 tienen una contribucion significativa positiva con el componente 10. Y negativas con 1 y 9.

En la grafica de PCA podemos ver que la varianza en la primera dimension de 37% y en la segunda de 17% tal que que la primera dimension principal captura mas informacion que la segunda, lo que sugiere que las variables originales estan mas alineadas. Podemos observar que muchas de las variables no se correlacionan de manera positiva entre ellas. La correlacion mas positiva es entre ProdElec y PNB95 mientras EmisCO2 y PropDefor tiene una correlacion altamente negativa.