Componentes Principales

Actividad 2

Luis Ángel Guzmán Iribe - A01741757

26 de Septiembre de 2023

Parte 1

A partir de los datos sobre indicadores económicos y sociales de 96 países 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.

```
# Paso 1
S = cov(M)
R = cor(M)

# Paso 2
eig_S = eigen(S)
eig_R = eigen(R)

# Paso 3
prop_var_S = eig_S$values / sum(diag(S))
prop_var_R = eig_R$values / sum(diag(R))

# Paso 4
cum_prop_var_S = cumsum(prop_var_S)
cum_prop_var_R = cumsum(prop_var_R)

cum_prop_var_S
```

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

```
cum_prop_var_R
```

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

Los arreglos de cum_prop_var_S y cum_prop_var_R me informan de la cantidad de la varianza explicada dependiendo de la cantidad de componentes principales empelados. Si observamos el resultado de la matriz de varianza-covarianza, podemos observar que podemos explicar el 0.9999 del modelo unicamente con los 2 primeros componentes principales, luego de los cuales estos se estavilizan y no aportan mucho a la explicación de la varianza.

Para la matriz de correlación, podemos apreciar que el primer compoennte principal explica el 0.36 de la varianza, y luego sube en intervalos disnimuyentes, necesitando la práctica totalidad de los componentes principales para explicar satisfactoriamente la varianza.

```
# Variables más influyentes por eigenvectors para los primeros 2 compoenntes principales
eigenvectors_S <- abs(eig_S$vectors[, 1:2])
```

```
colnames(df)
```

```
## NULL
```

```
eigenvectors_S[, 1]
```

```
##  [1] 1.658168e-06 4.048139e-05 5.739096e-06 8.880376e-01 4.597636e-01
##  [6] 3.504341e-04 2.625508e-04 4.089564e-06 1.073825e-06 2.547156e-03
## [11] 4.643724e-06
```

```
eigenvectors_S[, 2]
```

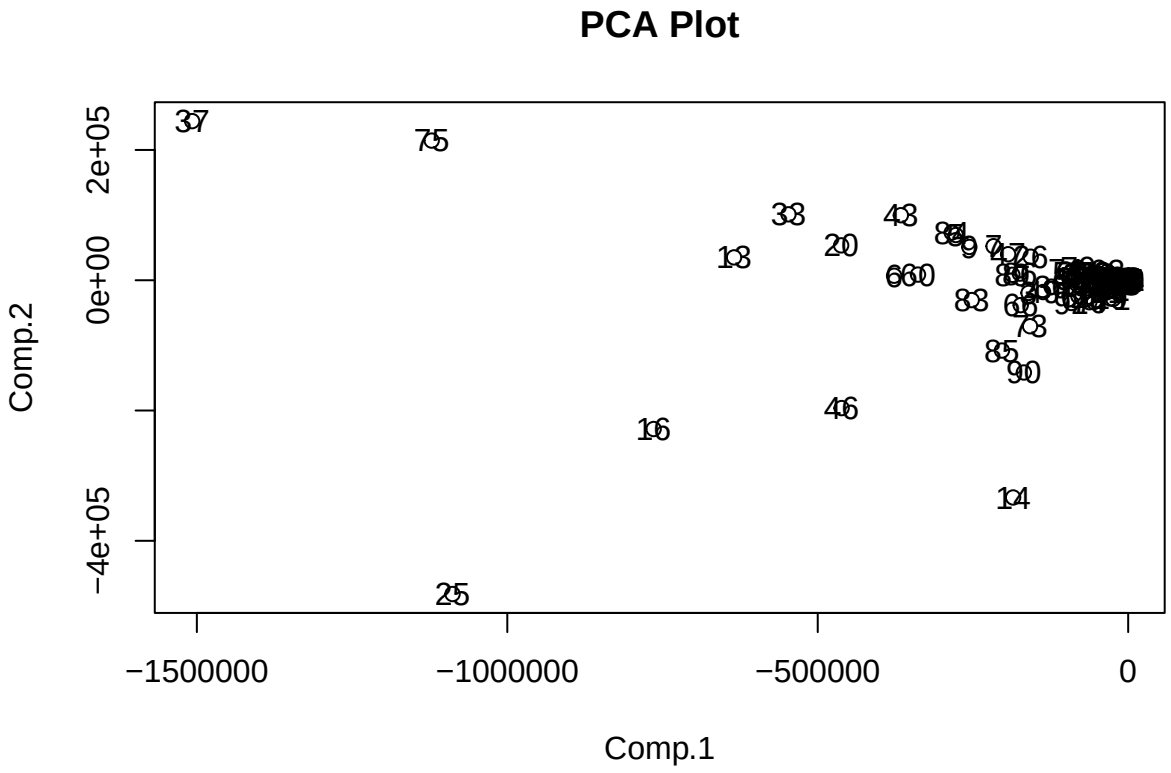
```
##  [1] 4.706785e-07 1.774254e-05 1.084543e-05 4.597632e-01 8.880405e-01
##  [6] 4.016179e-04 1.122118e-03 7.790843e-06 2.350808e-07 7.126782e-04
## [11] 1.315731e-06
```

Analizando los vectores del aporte de cada variable para los eigenvectors de los primeros 2 componentes principales, notamos que hay 2 variables que resaltan por ordenes de magnitud sobre los demás, 'ProdElec' y 'PNB85', lo que me indica a pensar que son estas variables las que tienen un mayor impacto sobre los componentes principales.

Parte 2

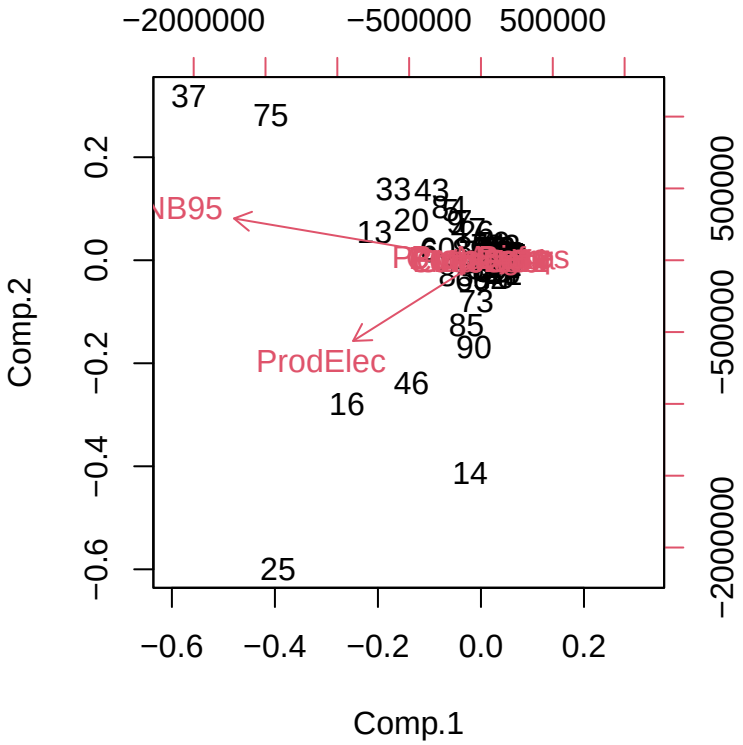
Obtenga las gráficas de respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes e interprete los resultados en término de agrupación de variables (puede ayudar "índice de riqueza", "índice de ruralidad")

```
cpS = princomp(M,cor=FALSE)
cpaS = as.matrix(M)%*%cpS$loadings
plot(cpaS[,1:2],type="p", main = "PCA Plot")
text(cpaS[,1],cpaS[,2],1:nrow(cpaS))
```



```r
biplot(cpS)
```

```
## 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
```

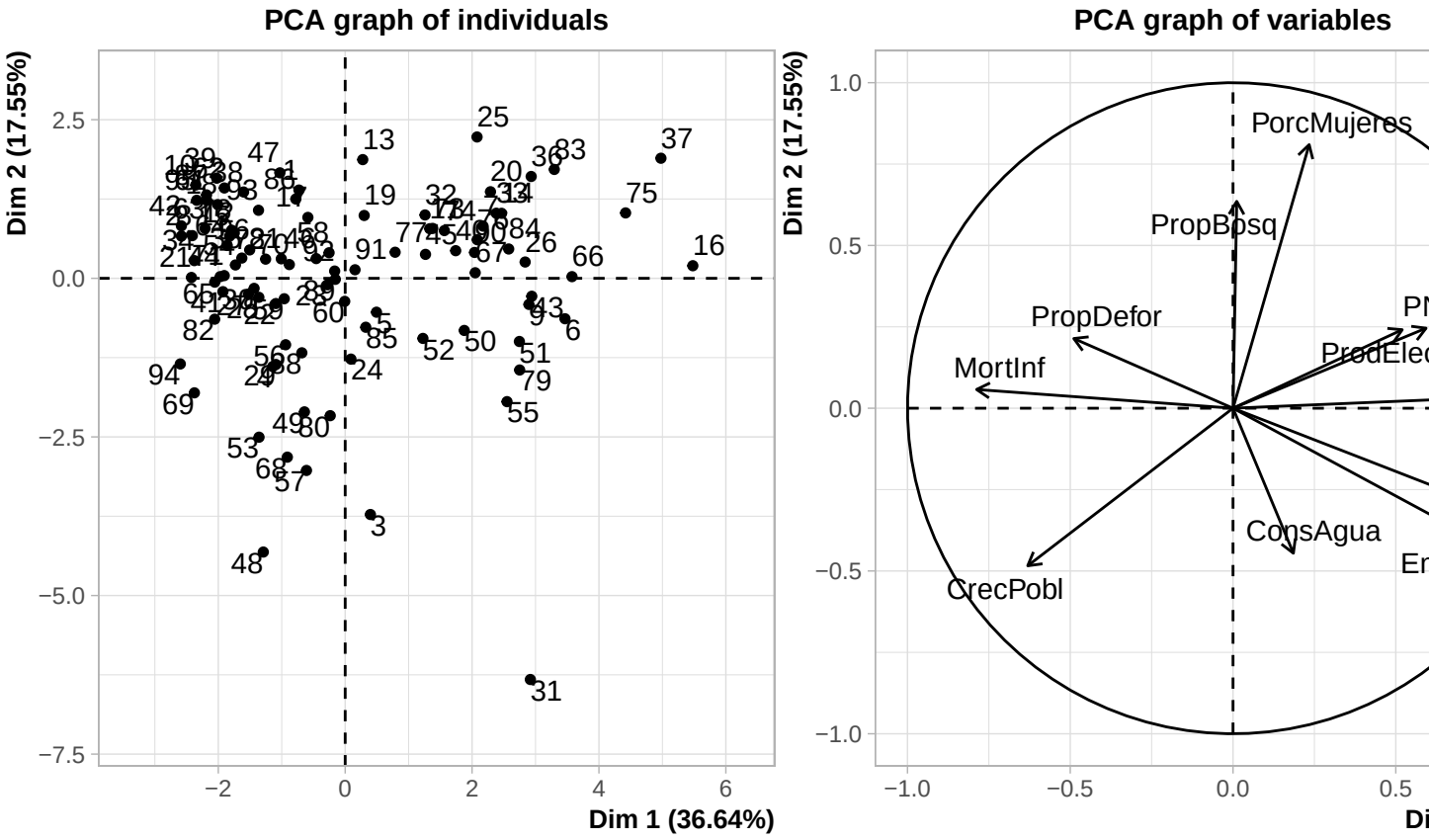


1. La primera gráfica nos muestra los datosen su gráfica de PCA con los componentes principales 1 y 2, como es de esperarse, la mayoría de la disperción se ecuentra a lo largo del primer componente principal, dado que es la que explica mayor porcentaje de varianza. 2. La segunda gráfica nos muestra los Eigenvectors para los componentes, entre los que destacan las ya mencionadas variables de 'ProdElec' y 'PNB85', las cuales comentavamos superan por ordenes de magnitud al resto de variables.

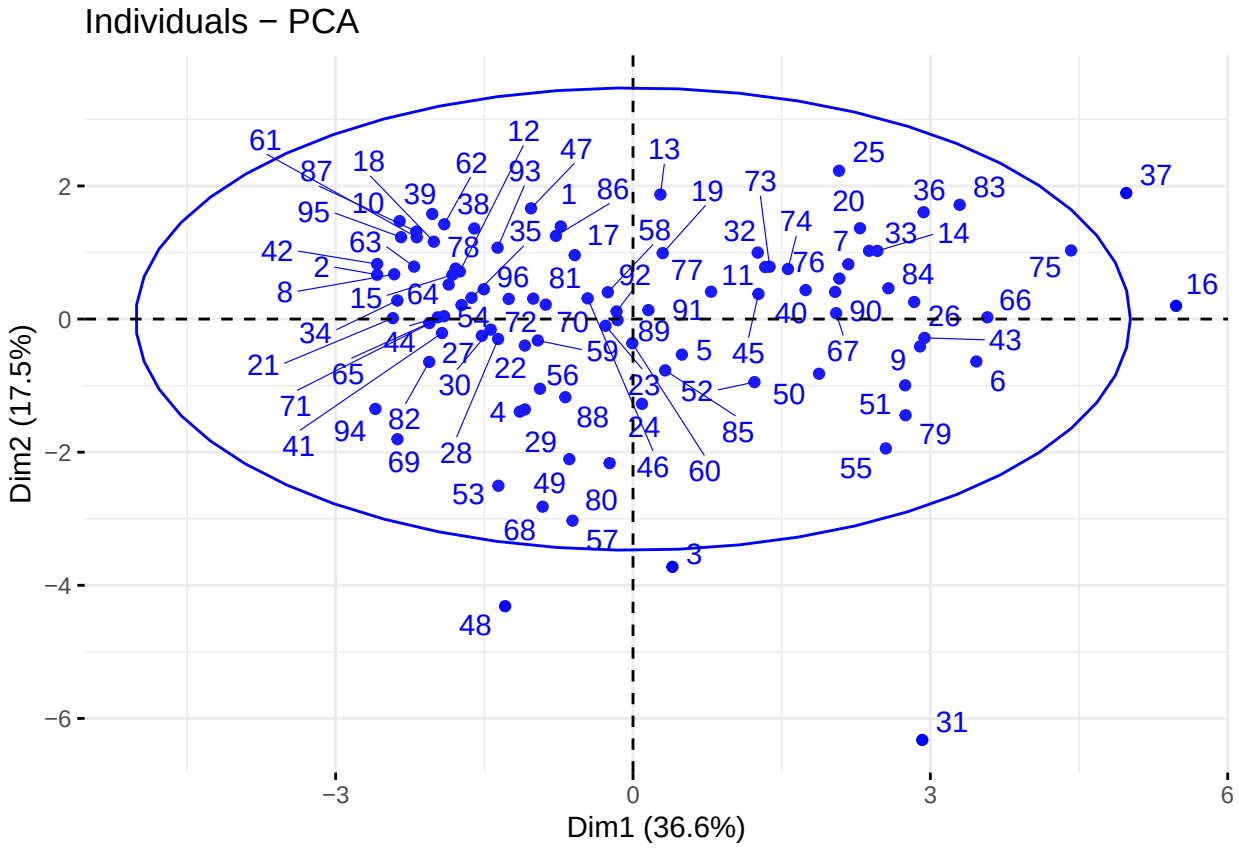
Parte 3

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

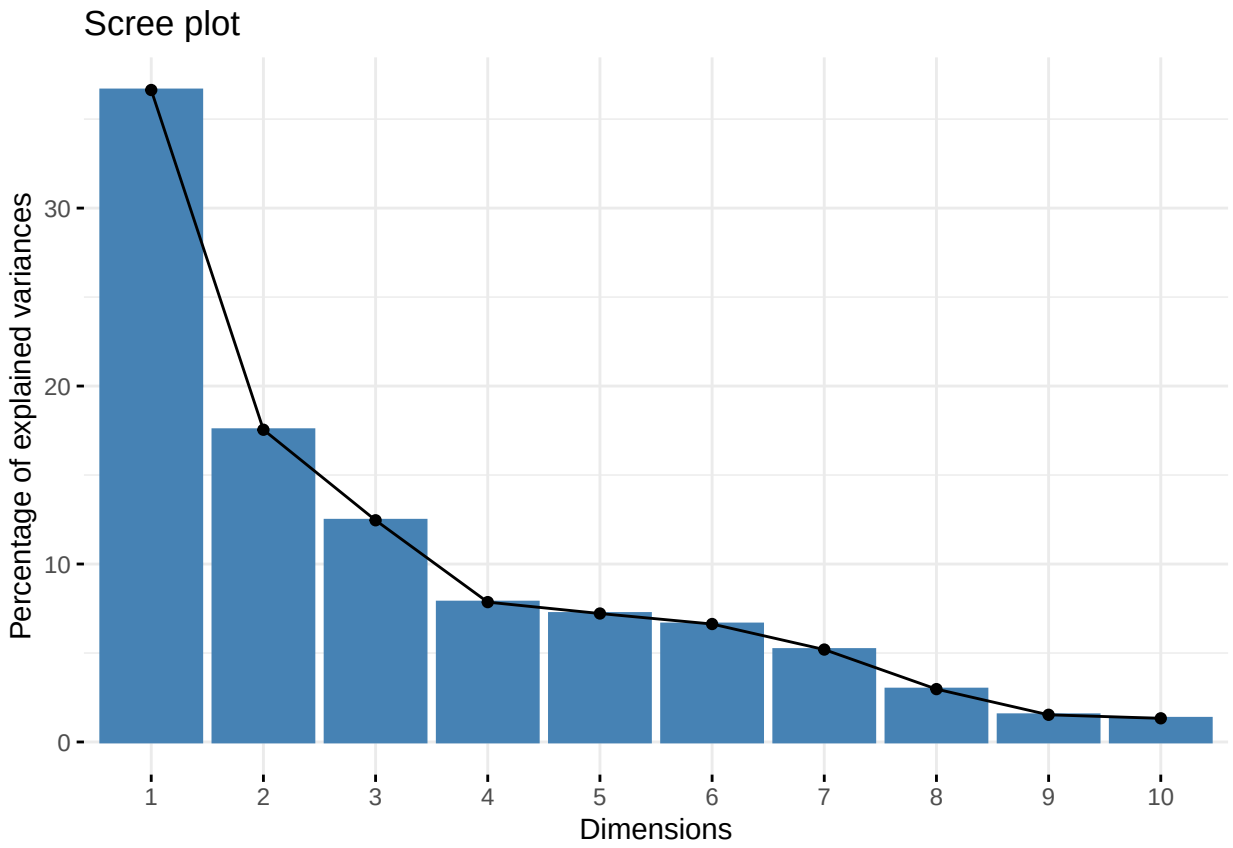
```
cp3 = PCA(M)
```



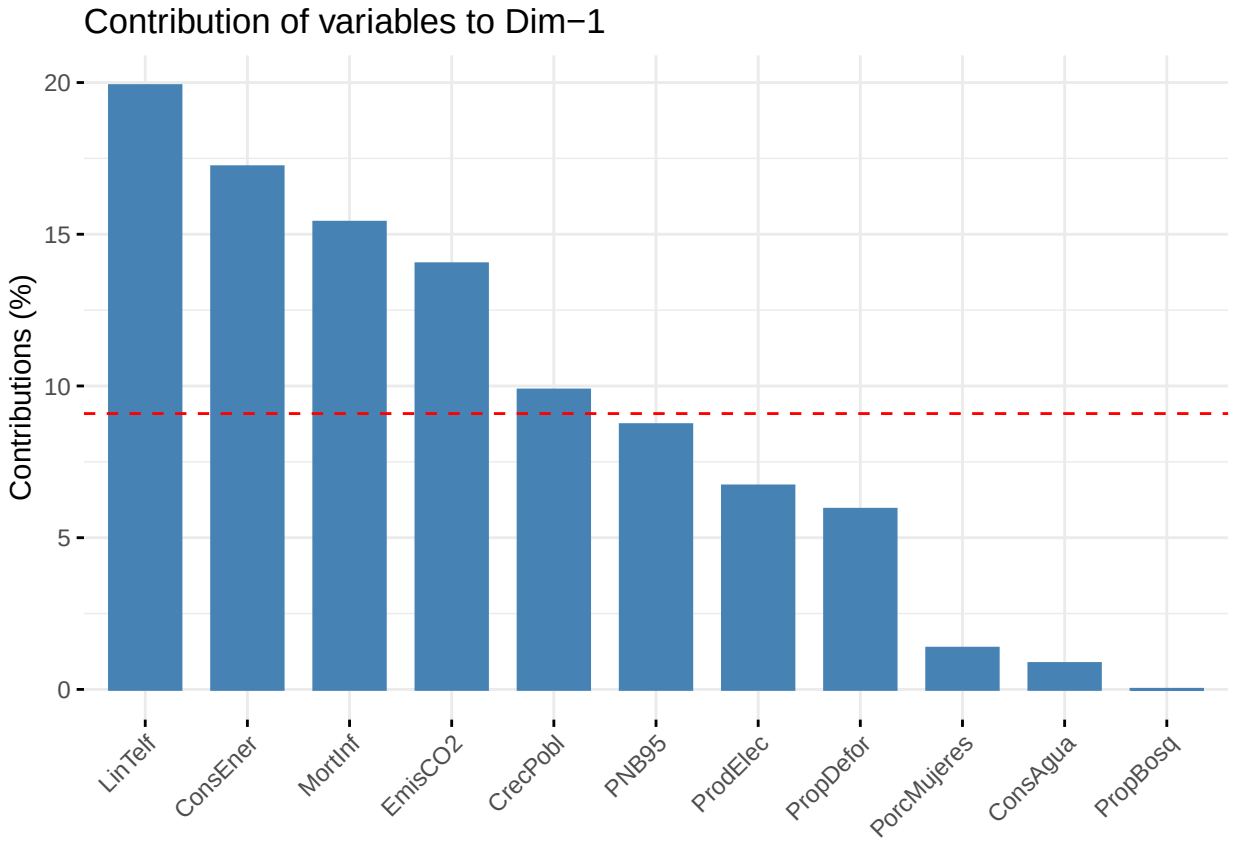

```
fviz_pca_ind(cp3, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




```
fviz_screeplot(cp3)
```




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



1. El gráfico, 'PCA graph of individuals' nos permite ver los individuos utilizados para el análisis, y la manera en la que estos se agrupan
2. El gráfico 'PCA of graph of variables' representa los eigenvectors para las variables del conjunto de datos, como podemos apreciar, ninguna es en exceso dominante sobre el resto, a diferencia de cuando utilizamos la matriz de varianzas-covarianzas.
3. La scree plot nos permite conocer el porcentaje de varianza proporcionado por cada componente principal, podemos apreciar los que los valores que más contribuyen son los primeros 3 amazando alrededor del 66% de la explicación de la varianza en conjunto. Ya habíamos notado en el primer punto de la tarea que la matriz de correlación, que es la que se utiliza para este ultimo anális, requiere de la práctica totalidad de los componentes principales para proveer de un gráfico satisfactorio.
4. La contribución de variables muestra precisamente eso, en que porcentaje las variables afecatn a los componentes principales. Mientras que la variable más significativa es LinTelf, no podemos afirmar que es dominante sobre el resto de variables, por lo que reducir los componentes del modelo utilizando esta tecnica sería complicado.