PCA

Diego Isau Barranco Herrera

Principal Component Analysis (PCA)

El PCA es una técnica de reducción de dimensión. A través de combinaciones lineales de las variables involucradas podemos reducir la cantidad de estas, siempre y cuando no se pierda información, esto es, tener como mínimo una varianza acumulada del 70% entre todos los componentes

Población USA

Tenemos 19 variables independientes, vamos a separar por año para realizar el PCA

```
poblacion <- read_xlsx("Covid.xlsm")
```

PCA para el año 2000

```
poblacion_2000 <- poblacion %>% select(c(2,3,5,7,9,11,13,15,17,19))

# normalizando datos
poblacion_2000_normal <- scale(poblacion_2000)

# Correlación
det(cor(poblacion_2000_normal))
```

```
[1] 1.359132e-40
```

La correlación tiende a cero, por lo que es adecuado

Ahora vamos a verificar el Factor de adecuación muestral de Kaiser, el cual nos indica si las variables son aptas para realizar el PCA

```r
psych::KMO(poblacion_2000_normal)
```

```
Error in solve.default(r) :
  system is computationally singular: reciprocal condition number = 1.00145e-18
```

```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = poblacion_2000_normal)
Overall MSA =  0.5
MSA for each item =
   Census Resident Total Population - AB:Qr-1-2000
                                                0.5
   Resident Total Population Estimate - Jul-1-2000
                                                0.5
              Net Domestic Migration - Jul-1-2000
                                                0.5
Federal/Civilian Movement from Abroad - Jul-1-2000
                                                0.5
          Net International Migration - Jul-1-2000
                                                0.5
                       Period Births - Jul-1-2000
                                                0.5
                       Period Deaths - Jul-1-2000
                                                0.5
Resident Under 65 Population Estimate - Jul-1-2000
                                                0.5
 Resident 65 Plus Population Estimate - Jul-1-2000
                                                0.5
                            Residual - Jul-1-2000
                                                0.5
```

Tenemos un MSA de 0.5, por lo que es una métrica útil y es pertinente hacer el PCA

Verificaremos cuántos componentes son adecuados para que tengamos mínimo una acumualción del 70% de la varianza

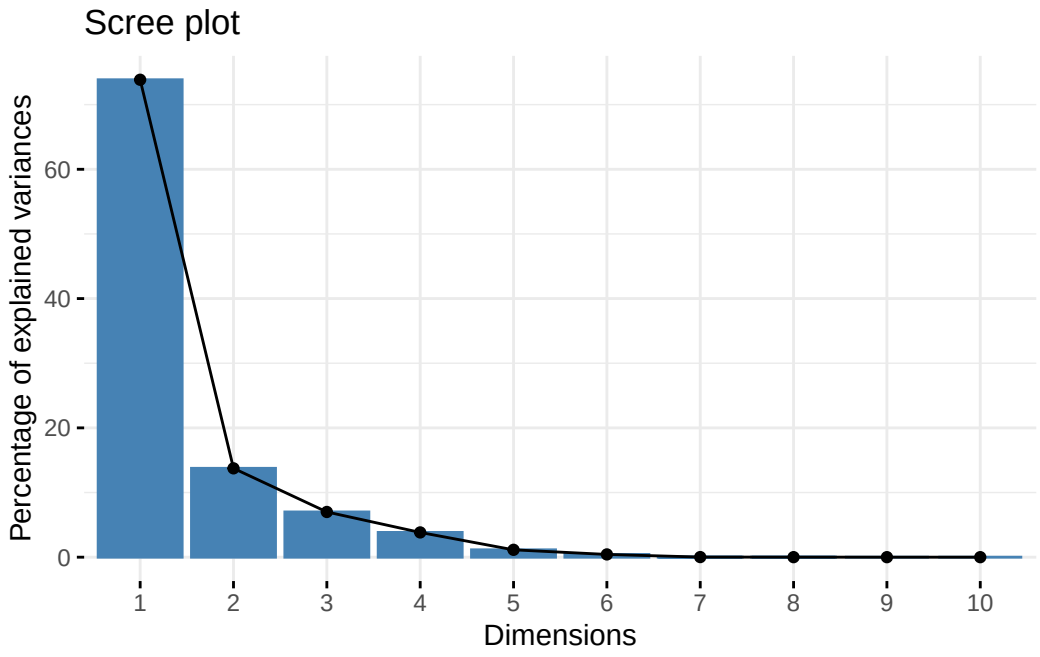
```r
pca_pob_2000 <- princomp(poblacion_2000_normal)
summary(pca_pob_2000)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     2.6907367 1.1607536 0.8280683 0.6125685 0.33421800
```

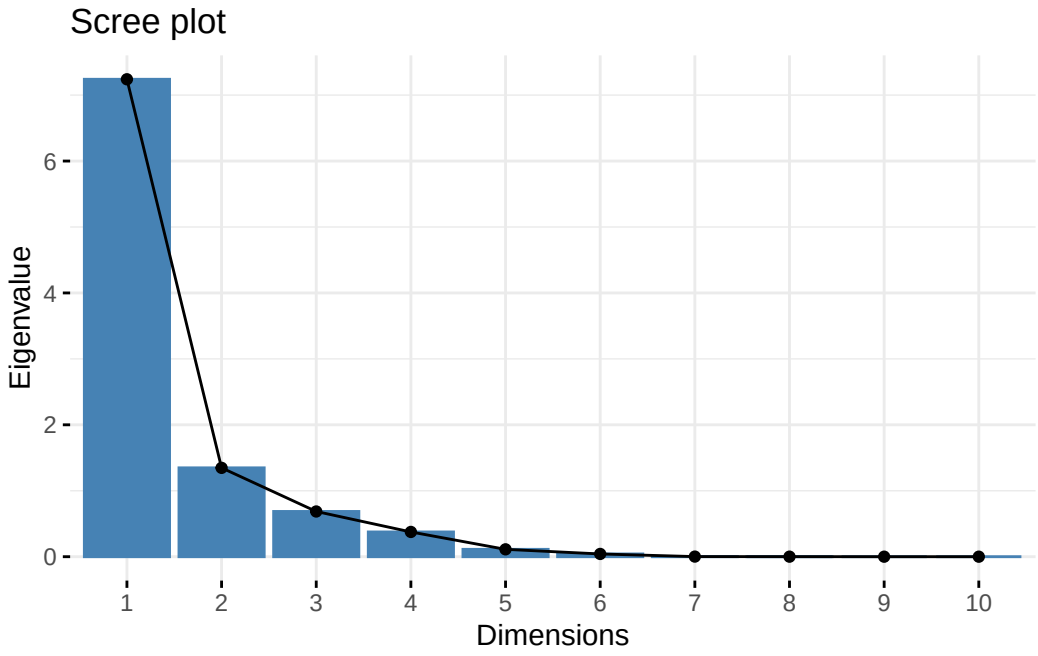
```
Proportion of Variance 0.7384865 0.1374296 0.0699411 0.0382745 0.01139357
Cumulative Proportion  0.7384865 0.8759161 0.9458572 0.9841317 0.99552529
                            Comp.6       Comp.7       Comp.8 Comp.9 Comp.10
Standard deviation     0.204077430 0.0391252568 2.629360e-02      0       0
Proportion of Variance 0.004248055 0.0001561401 7.051803e-05      0       0
Cumulative Proportion  0.999773342 0.9999294820 1.000000e+00      1       1
```

Gráficamente

```
fviz_eig(pca_pob_2000, choice='variance')
```



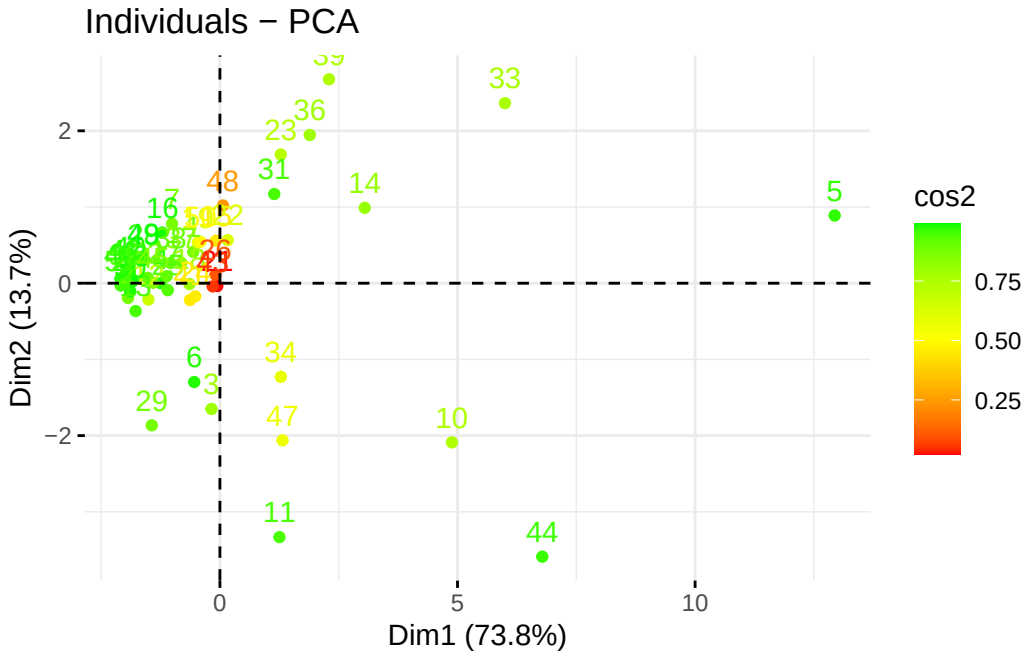

```
fviz_eig(pca_pob_2000, choice ='eigenvalue')
```



Con dos componentes se acumula el 87% de la varianza. Además se cumple que el eigenvalue de los dos componentes sea mayor a 1

Veamos cuánto contribuye cada variable a los diferentes componentes principales

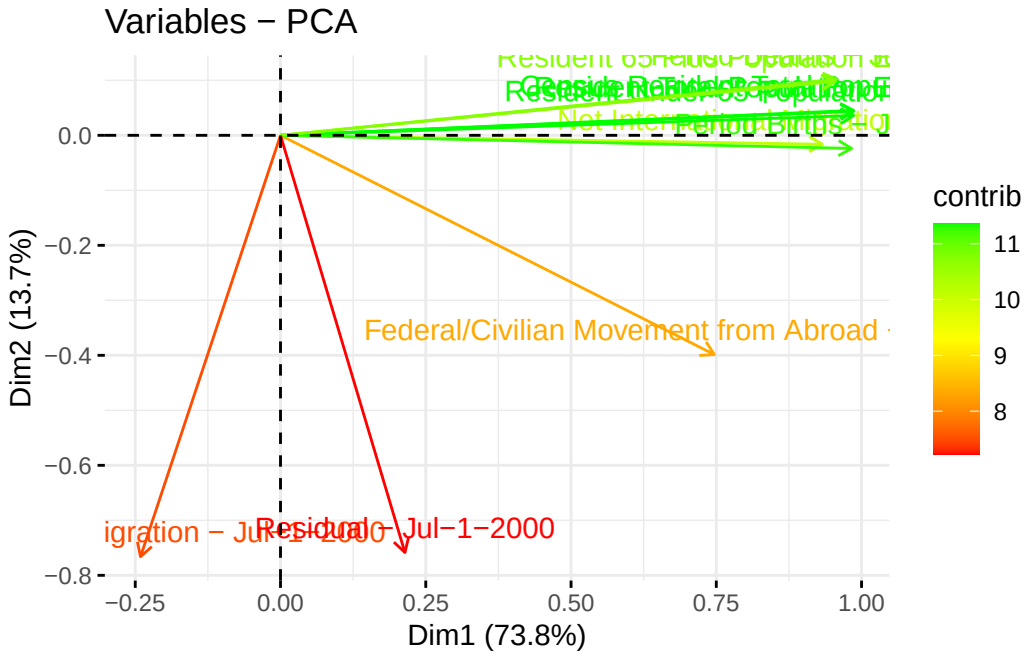
```
fviz_pca_ind(pca_pob_2000,
             col.ind = 'cos2',
gradient.cols = c('red', 'yellow', 'green'),
             repel = FALSE)
```



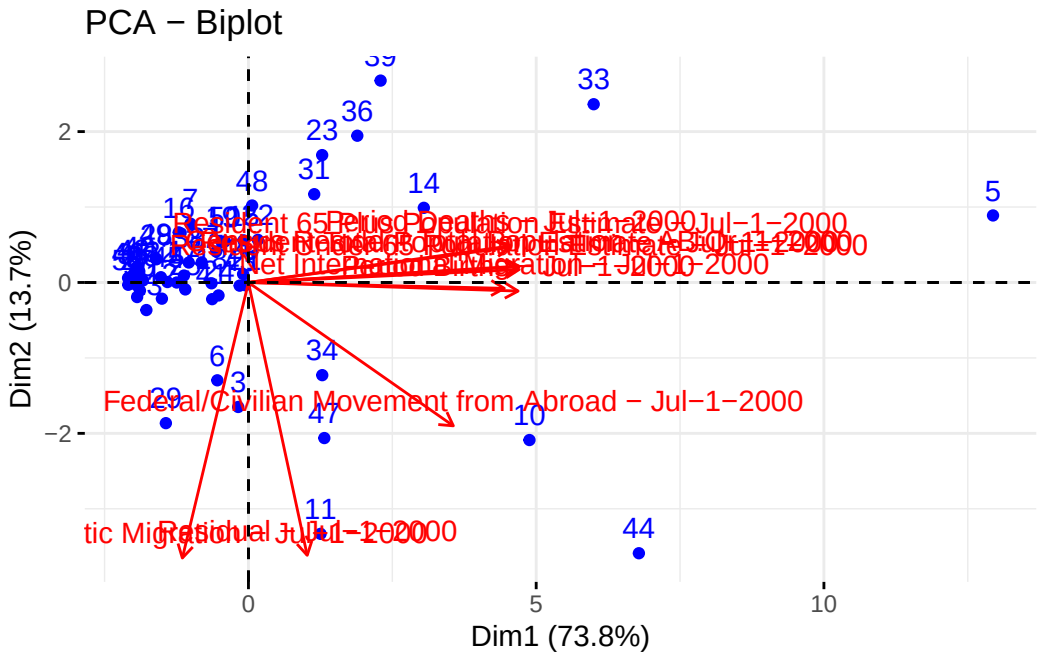
La mayoria de las observaciones se adecuan a dos dimensiones

Veamos cuánto contribuye cada variable a los diferentes componentes principales

```
fviz_pca_var(pca_pob_2000, col.var = 'contrib',
             gradient.cols = c('red', 'yellow', 'green'),
             repel = FALSE)
```




```
fviz_pca_biplot(pca_pob_2000, col.var='red', col.ind = 'blue')
```



Los componentes contribuyen en diferente medida a los cuadrantes dentro de la representación bidimensional

Haciendo el PCA con dos componentes

```
pca2 <- psych::principal(poblacion_2000_normal, nfactors=2,
                         residuals=FALSE, rotate="varimax",
                         scores=TRUE, oblique.scores=FALSE,
                         method='regression', use='pairwise',
                         cor='cor', weight=NULL)
```

```
Warning in cor.smooth(r): Matrix was not positive definite, smoothing was done
```

```
In factor.stats, I could not find the RMSEA upper bound . Sorry about that
```

```
Warning in psych::principal(poblacion_2000_normal, nfactors = 2, residuals =
FALSE, : The matrix is not positive semi-definite, scores found from Structure
loadings
```

```
pca2
```

```
Principal Components Analysis
Call: psych::principal(r = poblacion_2000_normal, nfactors = 2, residuals = FALSE,
    rotate = "varimax", scores = TRUE, oblique.scores = FALSE,
    method = "regression", use = "pairwise", cor = "cor", weight = NULL)
Standardized loadings (pattern matrix) based upon correlation matrix
                                                   RC1   RC2   h2     u2 com
Census Resident Total Population - AB:Qr-1-2000     1.00 -0.02 0.99 0.0059 1.0
Resident Total Population Estimate - Jul-1-2000     1.00 -0.02 0.99 0.0058 1.0
Net Domestic Migration - Jul-1-2000                -0.26  0.77 0.66 0.3421 1.2
Federal/Civilian Movement from Abroad - Jul-1-2000  0.74  0.42 0.73 0.2692 1.6
Net International Migration - Jul-1-2000            0.94  0.04 0.89 0.1128 1.0
Period Births - Jul-1-2000                          0.99  0.05 0.99 0.0142 1.0
Period Deaths - Jul-1-2000                          0.97 -0.08 0.94 0.0563 1.0
Resident Under 65 Population Estimate - Jul-1-2000  1.00 -0.01 0.99 0.0061 1.0
Resident 65 Plus Population Estimate - Jul-1-2000   0.97 -0.07 0.94 0.0623 1.0
Residual - Jul-1-2000                               0.20  0.77 0.63 0.3663 1.1

                       RC1  RC2
SS loadings           7.38 1.38
Proportion Var        0.74 0.14
Cumulative Var        0.74 0.88
Proportion Explained  0.84 0.16
```

```
Cumulative Proportion 0.84 1.00

Mean item complexity =  1.1
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  15.64  with prob <  0.94

Fit based upon off diagonal values = 0.99
```

Dos componentes son suficientes para explicar la mayor parte de la variabilidad (**88%**), con un ajuste estadísticamente sólido. El primer componente domina en importancia, pero el segundo aporta información complementaria. Este modelo es válido para reducir la dimensionalidad de los datos sin perder información crítica.

PCA para el año 2001

Se realzará el mismo procecimiento que en el año 2000 y al final se darán las interpretaciones

```
poblacion_2001 <- poblacion %>% select(c(4,6,8,10,12,14,16,18,20))

# normalizando datos
poblacion_2001_normal <- scale(poblacion_2001)

# Correlación
det(cor(poblacion_2001_normal))  # La correlación tiende a cero
```

```
[1] -3.985373e-26
```

```
#psych::cor.plot(poblacion_2000_normal)
```

La correlación tiende a cero, por lo que es adecuado

Ahora vamos a verificar el Factor de adecuación muestral de Kaiser

```
psych::KMO(poblacion_2001_normal)
```

```
Error in solve.default(r) :
  system is computationally singular: reciprocal condition number = 5.86373e-18
```

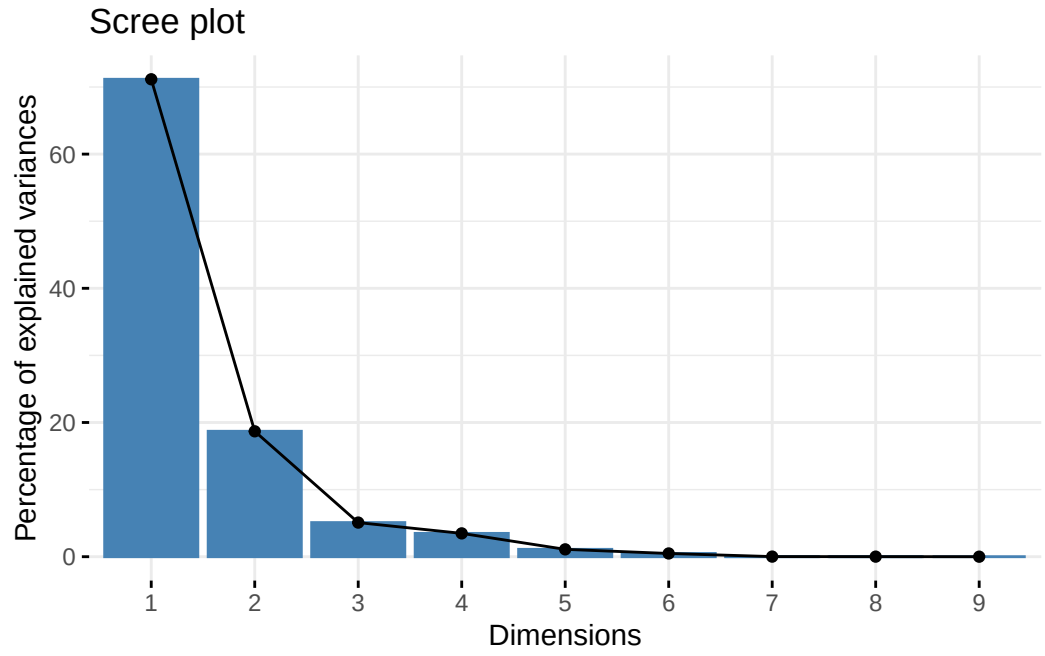
```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = poblacion_2001_normal)
Overall MSA =  0.5
MSA for each item =
   Resident Total Population Estimate - Jul-1-2001
                                                0.5
              Net Domestic Migration - Jul-1-2001
                                                0.5
Federal/Civilian Movement from Abroad - Jul-1-2001
                                                0.5
          Net International Migration - Jul-1-2001
                                                0.5
                       Period Births - Jul-1-2001
                                                0.5
                       Period Deaths - Jul-1-2001
                                                0.5
Resident Under 65 Population Estimate - Jul-1-2001
                                                0.5
 Resident 65 Plus Population Estimate - Jul-1-2001
                                                0.5
                             Residual - Jul-1-2001
                                                0.5
```

```r
pca_pob_2001 <- princomp(poblacion_2001_normal)
summary(pca_pob_2001)
```

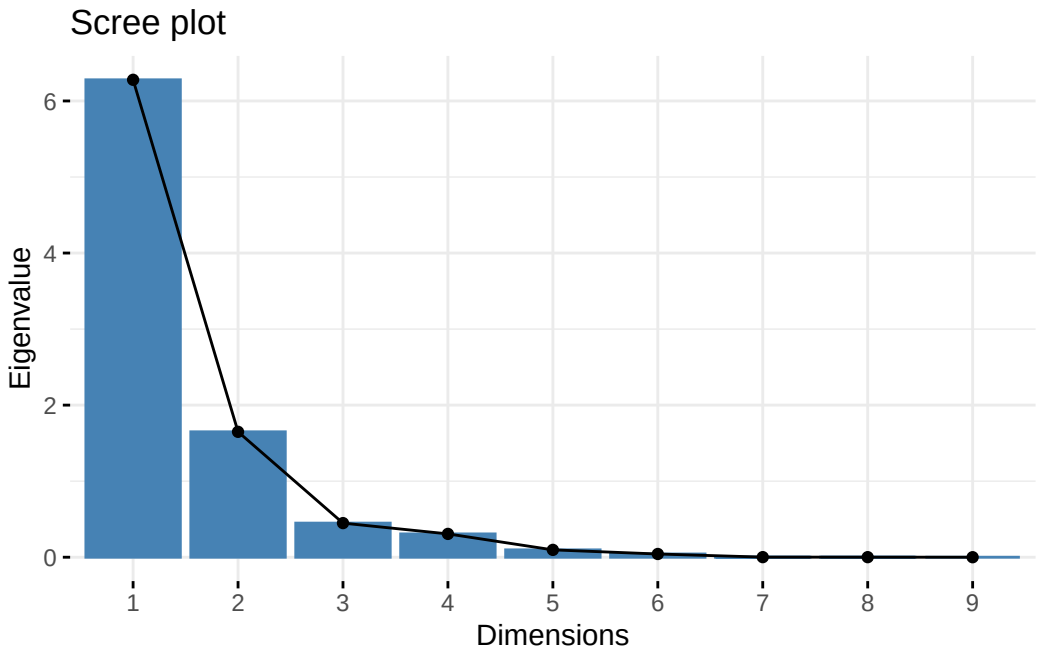
```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.5056381 1.2841609 0.66974585 0.55374490 0.31095905
Proportion of Variance 0.7115319 0.1868945 0.05083674 0.03475179 0.01095883
Cumulative Proportion  0.7115319 0.8984264 0.94926311 0.98401490 0.99497372
                           Comp.6       Comp.7       Comp.8 Comp.9
Standard deviation     0.205937451 0.0345666035 2.728378e-02      0
Proportion of Variance 0.004806493 0.0001354163 8.436586e-05      0
Cumulative Proportion  0.999780218 0.9999156341 1.000000e+00      1
```

Gráficamente

```r
fviz_eig(pca_pob_2001, choice='variance')
```

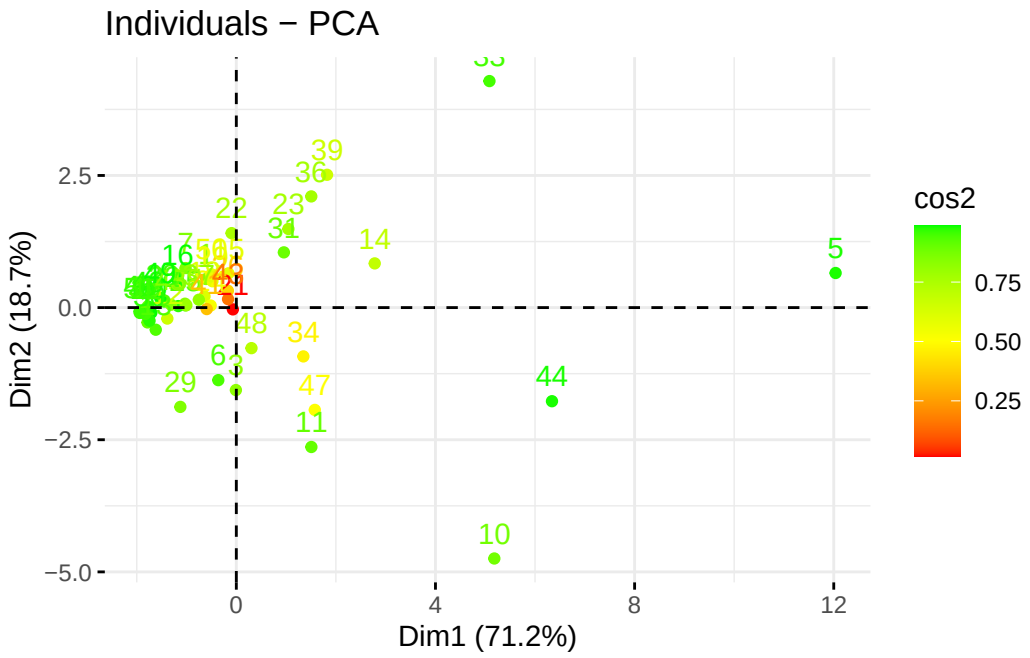



```
fviz_eig(pca_pob_2001, choice ='eigenvalue')
```



Con dos componentes se acumula el 89% de la varianza. Además se cumple que el eigenvalue de los dos componentes sea mayor a 1

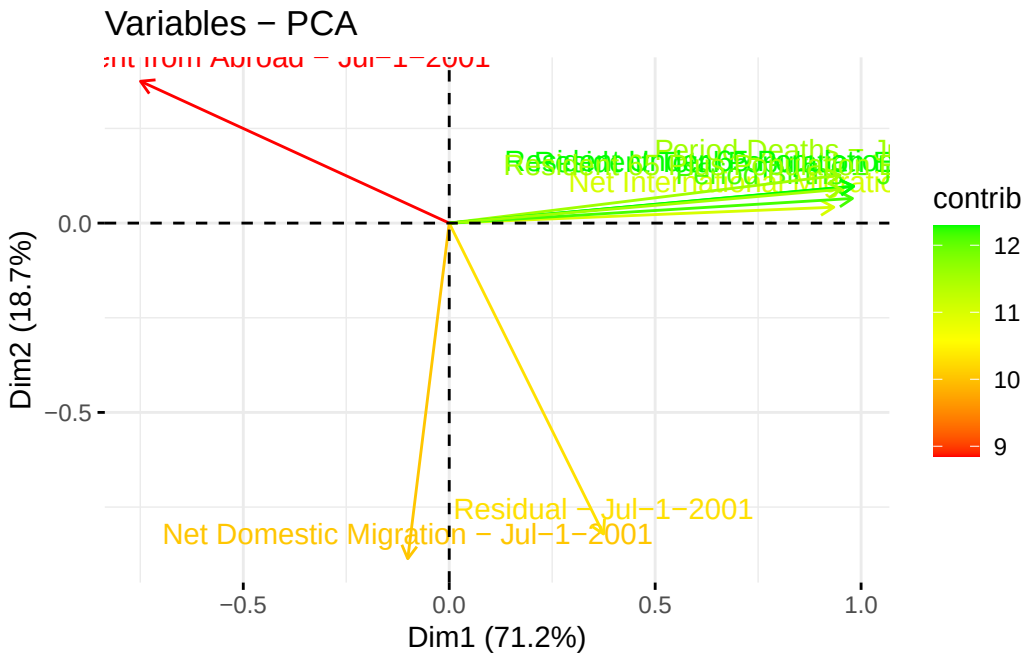
```
# Grafico de las puntuaciones factoriales y su representación
fviz_pca_ind(pca_pob_2001,
             col.ind = 'cos2',gradient.cols = c('red', 'yellow', 'green'),
             repel = FALSE)
```



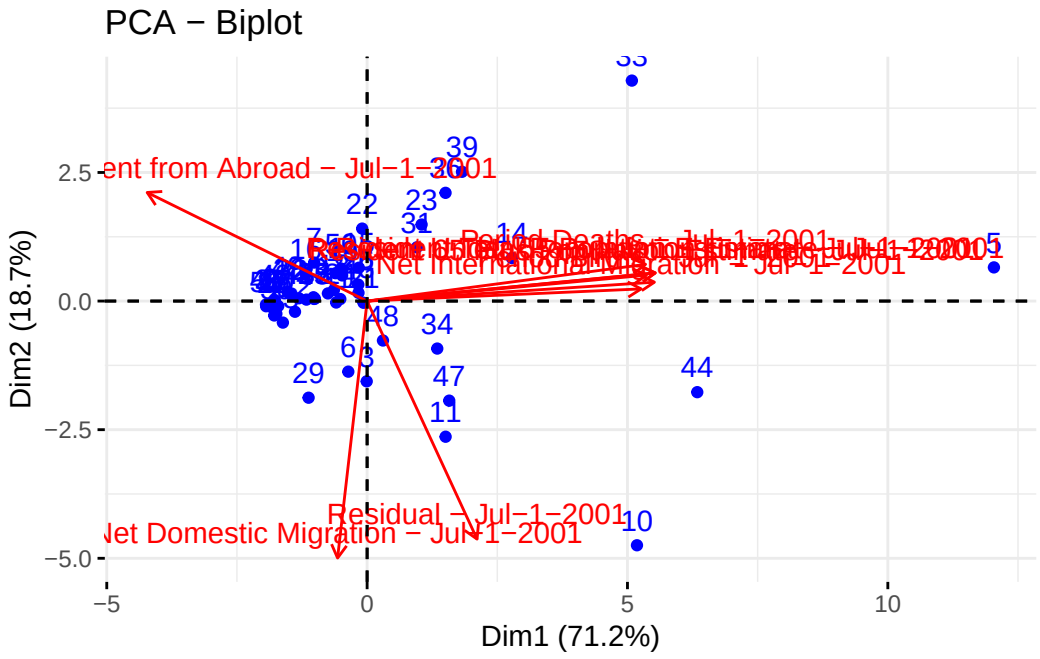
La mayoria de las observaciones se adecuan a dos dimensiones

Veamos cuánto contribuye cada variable a los diferentes componentes principales

```
# cuanto contribuye a cada variable
fviz_pca_var(pca_pob_2001, col.var = 'contrib',
             gradient.cols = c('red', 'yellow', 'green'),
             repel = FALSE)
```




```
fviz_pca_biplot(pca_pob_2001, col.var='red', col.ind = 'blue')
```



Ahora haremos el PCA con dos componentes

```r
pca2 <- psych::principal(poblacion_2001_normal, nfactors=2,
                         residuals=FALSE, rotate="varimax",
                         scores=TRUE, oblique.scores=FALSE,
                         method='regression', use='pairwise',
                         cor='cor', weight=NULL)
```

```
Warning in cor.smooth(r): Matrix was not positive definite, smoothing was done
```

```
Warning in psych::principal(poblacion_2001_normal, nfactors = 2, residuals =
FALSE, : The matrix is not positive semi-definite, scores found from Structure
loadings
```

```r
pca2
```

```
Principal Components Analysis
Call: psych::principal(r = poblacion_2001_normal, nfactors = 2, residuals = FALSE,
    rotate = "varimax", scores = TRUE, oblique.scores = FALSE,
    method = "regression", use = "pairwise", cor = "cor", weight = NULL)
Standardized loadings (pattern matrix) based upon correlation matrix
                                                    RC1   RC2   h2     u2 com
Resident Total Population Estimate - Jul-1-2001    1.00  0.03 0.99 0.0068 1.0
Net Domestic Migration - Jul-1-2001               -0.22  0.87 0.81 0.1884 1.1
Federal/Civilian Movement from Abroad - Jul-1-2001 -0.70 -0.47 0.72 0.2839 1.8
Net International Migration - Jul-1-2001           0.94  0.08 0.89 0.1088 1.0
Period Births - Jul-1-2001                         0.99  0.06 0.98 0.0178 1.0
Period Deaths - Jul-1-2001                         0.97  0.00 0.93 0.0651 1.0
Resident Under 65 Population Estimate - Jul-1-2001 1.00  0.03 0.99 0.0068 1.0
Resident 65 Plus Population Estimate - Jul-1-2001  0.96  0.04 0.93 0.0676 1.0
Residual - Jul-1-2001                              0.27  0.87 0.83 0.1691 1.2

                       RC1  RC2
SS loadings           6.32 1.76
Proportion Var        0.70 0.20
Cumulative Var        0.70 0.90
Proportion Explained  0.78 0.22
Cumulative Proportion 0.78 1.00

Mean item complexity =  1.1
Test of the hypothesis that 2 components are sufficient.
```

```
The root mean square of the residuals (RMSR) is  0.05
 with the empirical chi square  7.81  with prob <  0.99

Fit based upon off diagonal values = 1
```

Igualemnte, dos componentes son altamente suficientes, explicando el **90%** de la varianza con un ajuste estadístico muy bueno. El segundo componente gana relevancia en 2001, lo que podría indicar cambios en la estructura subyacente de los datos o una mejor captura de variables. El modelo es óptimo para reducir dimensionalidad sin pérdida crítica de información.

Comparando los años

- Mayor varianza acumulada (90% vs 88%) y mejor distribución entre componentes (RC2 explica 20% vs 14% en 2000).
- Estadísticos de ajuste mejorados (RMSR más bajo, chi-cuadrado menor y p-valor más alto).

data_pca

```
data_pca <- read.csv2("data_pca.csv")

# normalizar datos
data_normal <- scale(data_pca[,-16])
```

Veamos las estadísticas básicas para determinar si es viable realizar el PCA

```
# determinante de correlación
det(cor(data_normal))  # La correlación tiende a cero
```

```
[1] 0.004667778
```

```
# Factor de adecuación muestral de kaiser
psych::KMO(data_normal)  #
```

```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = data_normal)
Overall MSA =  0.34
MSA for each item =
  x1   x2   x3   x4   x5   x6   x7   x8   x9  x10  x11  x12  x13  x14  x15
0.36 0.27 0.24 0.46 0.53 0.55 0.45 0.27 0.42 0.26 0.43 0.46 0.28 0.62 0.33
```

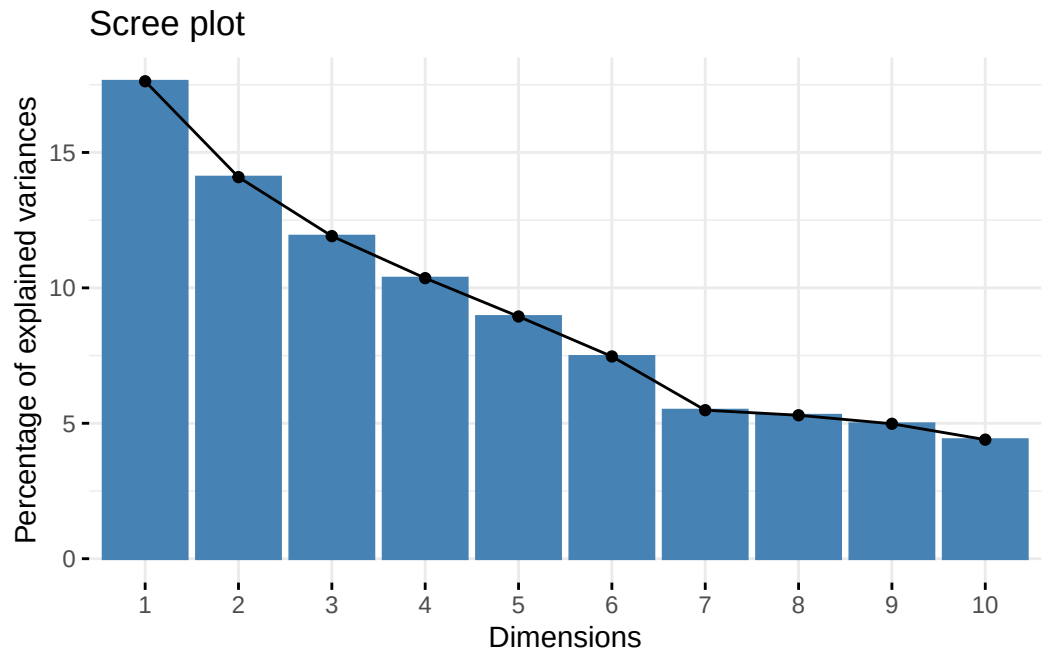
Tenemos un MSA de 0.34, es muy bajo por lo que no se recomienda PCA, sin embargo procedemos con el ejercicio

```
pca <- princomp(data_normal)
summary(pca)
```

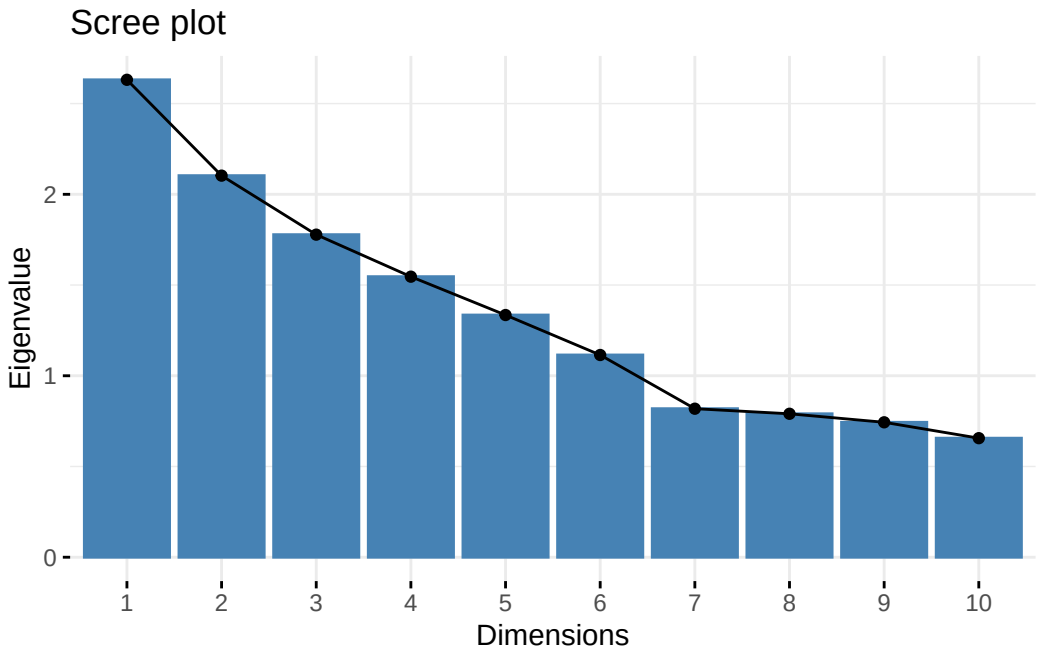
```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.6220588 1.4501268 1.3332930 1.2434264 1.15529908
Proportion of Variance 0.1762864 0.1408957 0.1191069 0.1035919 0.08942821
Cumulative Proportion  0.1762864 0.3171821 0.4362890 0.5398809 0.62930907
                           Comp.6     Comp.7     Comp.8    Comp.9    Comp.10
Standard deviation     1.05569426 0.90471763 0.88908929 0.8622762 0.80999883
Proportion of Variance 0.07467272 0.05484181 0.05296347 0.0498171 0.04395967
Cumulative Proportion  0.70398179 0.75882360 0.81178707 0.8616042 0.90556384
                          Comp.11    Comp.12    Comp.13   Comp.14     Comp.15
Standard deviation     0.7012045 0.59518243 0.53958339 0.4662240 0.234552581
Proportion of Variance 0.0329439 0.02373482 0.01950755 0.0145638 0.003686091
Cumulative Proportion  0.9385077 0.96224255 0.98175010 0.9963139 1.000000000
```

Vamos a considerar 6 componentes, donde hay una acumulación de varianza del 70%, como se puede apreciar en las siguientes gráficas. A el igual que se hasta el componente 6 el eigenvalue es mayor a 1

```
# grafica de eigenvalores y varianza
fviz_eig(pca, choice='variance')
```

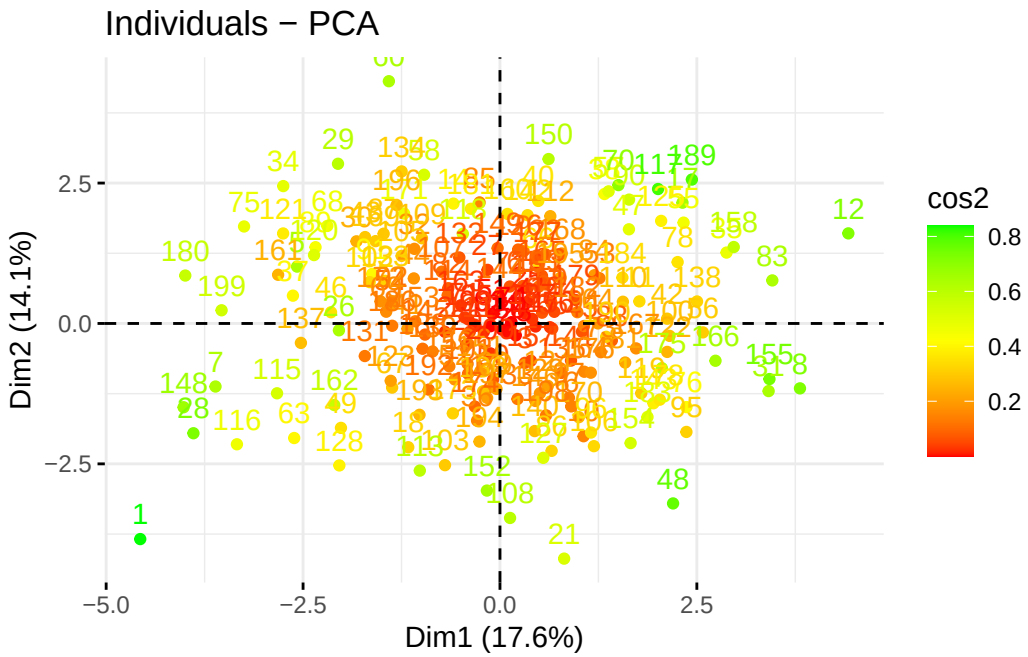



```
fviz_eig(pca, choice ='eigenvalue')
```



Sus puntuaciones factoriales

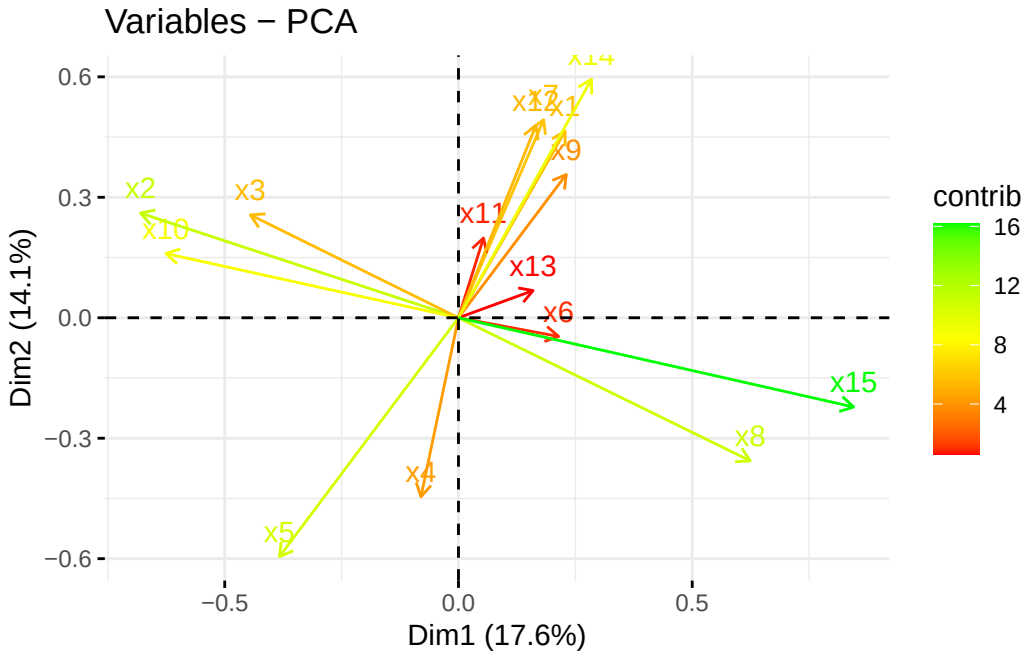
```
# Grafico de las puntuaciones factoriales y su representación
fviz_pca_ind(pca,
             col.ind = 'cos2',
             gradient.cols = c('red', 'yellow', 'green'),
             repel = FALSE)
```



Existe una gran cantidad de datos no bien representados correctamente

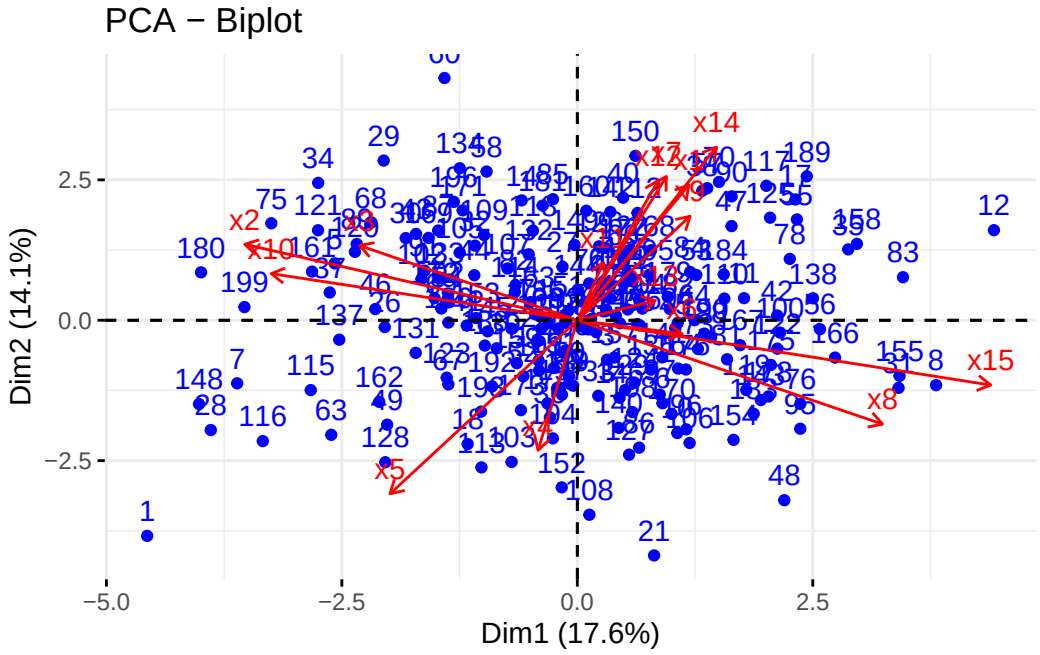
Veamos las cargas

```
fviz_pca_var(pca, col.var = 'contrib',
             gradient.cols = c('red', 'yellow', 'green'),
             repel = FALSE)
```



Las variables cercanas a los extremos de los ejes (RC1 o RC2) son las más influyentes en ese componente. Por ejemplo, una variable en la esquina superior derecha contribuye fuertemente a RC1.

```
fviz_pca_biplot(pca, col.var='red', col.ind = 'blue')
```



Respecto a las variables, su dirección y longitud indican su influencia en los componentes. Si una variable apunta hacia un grupo de observaciones, esas observaciones tienen valores altos en esa variable.

Realizamos el PCA con el número de componentes seguridos

```
# PCA con 6 componentes

pca_6comp <- psych::principal(data_normal, nfactors = 6,
                              residuals=FALSE, rotate="varimax",
                              scores=TRUE, oblique.scores=FALSE,
                              method='regression', use='pairwise',
                              cor='cor', weight=NULL)
pca_6comp
```

```
Principal Components Analysis
Call: psych::principal(r = data_normal, nfactors = 6, residuals = FALSE,
    rotate = "varimax", scores = TRUE, oblique.scores = FALSE,
    method = "regression", use = "pairwise", cor = "cor", weight = NULL)
Standardized loadings (pattern matrix) based upon correlation matrix
      RC1   RC2   RC4   RC3   RC6   RC5   h2   u2 com
x1  -0.06  0.57 -0.04  0.06 -0.10 -0.72 0.86 0.14 2.0
x2   0.80 -0.05  0.07  0.02  0.00 -0.15 0.68 0.32 1.1
x3   0.22  0.08 -0.84 -0.08  0.06  0.01 0.77 0.23 1.2
x4  -0.01  0.06  0.07 -0.50 -0.65  0.12 0.69 0.31 2.0
x5   0.10 -0.85  0.03 -0.03 -0.06 -0.07 0.75 0.25 1.1
x6   0.12  0.07  0.87  0.00  0.04  0.03 0.78 0.22 1.1
x7   0.02  0.03  0.05  0.87  0.21 -0.01 0.80 0.20 1.1
x8  -0.73 -0.13  0.12  0.11 -0.12  0.00 0.59 0.41 1.2
x9  -0.06  0.10  0.04  0.77 -0.12  0.00 0.63 0.37 1.1
x10  0.76 -0.07  0.13  0.02 -0.08  0.13 0.62 0.38 1.2
x11 -0.02 -0.07  0.03 -0.04  0.78  0.02 0.61 0.39 1.0
x12  0.02  0.41  0.00 -0.01  0.54  0.06 0.46 0.54 1.9
x13 -0.05  0.31  0.00 -0.01 -0.05  0.88 0.87 0.13 1.3
x14  0.03  0.80  0.04  0.09 -0.03  0.04 0.66 0.34 1.0
x15 -0.77  0.15  0.42 -0.03  0.07  0.00 0.80 0.20 1.7

                       RC1  RC2  RC4  RC3  RC6  RC5
SS loadings           2.42 2.04 1.69 1.63 1.42 1.35
Proportion Var        0.16 0.14 0.11 0.11 0.09 0.09
Cumulative Var        0.16 0.30 0.41 0.52 0.61 0.70
Proportion Explained  0.23 0.19 0.16 0.15 0.13 0.13
Cumulative Proportion 0.23 0.42 0.58 0.74 0.87 1.00
```

```
Mean item complexity =  1.3
Test of the hypothesis that 6 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  263.19  with prob <  4.2e-39

Fit based upon off diagonal values = 0.83
```

Los 6 componentes capturan **70%** de la varianza total, significativamente menor que el**88-90%** logrado con solo 2 componentes en análisis previos. Esto sugiere pérdida de eficiencia al incluir más componentes.

El p-value rechaza fuertemente la hipótesis de que 6 componentes son suficientes. Los residuos no son aleatorios, sugiriendo que el modelo no captura toda la estructura subyacente

> **!** Important
>
> [Link de GitHub](https://github.com/Diego195200/CC_course_2025A/tree/main/R%20code)