Práctica 3. Análisis de Componentes Principales

Santiago de Jesús Fuentes Hinojosa

Código completo

> **!** Link al Github
>
> Si se requiere visualizar el código completo, entra al link de mi Github:

Documento: DATA_PCA

Se realizará un Análisis de Componentes Principales sobre un conjunto de datos específico. Se evaluarán las principales direcciones de variabilidad de las observaciones y se examinarán las relaciones entre las variables del conjunto de datos. Este análisis nos permitirá una mejor comprensión de la estructura de los datos y proporcionará una base sólida para realizar análisis más profundos.

Código

Realizamos el código que explique el proceso que se siguió para la obtención de los componentes

```
require(tidyverse)

datos <- read.csv2("data_pca2.csv")
```

```
# Vemos como se comportan los datos
summary(datos)
```

```
       x1                x2                x3                x4          
 Min.   :-5.370   Min.   :-3.7700   Min.   :-4.2600   Min.   :-8.910  
 1st Qu.: 2.735   1st Qu.:-0.9650   1st Qu.:-0.1375   1st Qu.:-0.665  
 Median : 6.265   Median : 0.0000   Median : 1.6750   Median : 1.580  
 Mean   : 6.193   Mean   : 0.1887   Mean   : 1.6715   Mean   : 1.780  
 3rd Qu.: 9.870   3rd Qu.: 1.3625   3rd Qu.: 3.3050   3rd Qu.: 4.175  
 Max.   :18.620   Max.   : 5.4500   Max.   : 8.9300   Max.   :14.030  
       x5                x6                x7                x8          
 Min.   :-7.210   Min.   :-2.450   Min.   :-6.7000   Min.   :-3.9800  
 1st Qu.:-0.880   1st Qu.: 0.695   1st Qu.:-0.1525   1st Qu.: 0.2625  
 Median : 1.425   Median : 1.620   Median : 1.7850   Median : 1.5600  
 Mean   : 1.652   Mean   : 1.896   Mean   : 1.9162   Mean   : 1.6114  
 3rd Qu.: 4.220   3rd Qu.: 2.935   3rd Qu.: 4.2450   3rd Qu.: 3.2500  
 Max.   :10.650   Max.   : 6.580   Max.   :14.7100   Max.   : 6.9400  
       x9               x10               x11               x12          
 Min.   :-3.240   Min.   :-9.2200   Min.   :-2.8700   Min.   :-4.680  
 1st Qu.: 0.705   1st Qu.:-3.0625   1st Qu.:-0.5500   1st Qu.: 1.427  
 Median : 1.980   Median :-0.3950   Median : 0.2100   Median : 3.045  
 Mean   : 1.873   Mean   :-0.5763   Mean   : 0.2117   Mean   : 3.388  
 3rd Qu.: 2.928   3rd Qu.: 1.8100   3rd Qu.: 0.9800   3rd Qu.: 5.460  
 Max.   : 8.150   Max.   : 8.5700   Max.   : 2.9500   Max.   :10.790  
      x13               x14               x15                 y          
 Min.   :-9.8800   Min.   :-15.3700   Min.   :-6.91000   Min.   :-0.4700  
 1st Qu.:-3.2600   1st Qu.: -3.6075   1st Qu.:-1.86750   1st Qu.: 0.9575  
 Median :-1.2000   Median :  0.0750   Median : 0.13000   Median : 1.4200  
 Mean   :-1.1153   Mean   : -0.0944   Mean   :-0.04865   Mean   : 1.3754  
 3rd Qu.: 0.9325   3rd Qu.:  3.1950   3rd Qu.: 1.24000   3rd Qu.: 1.8300  
 Max.   :10.9200   Max.   : 13.7500   Max.   : 6.89000   Max.   : 3.0800  
```

> ⚠ Variable y
>
> Contamos con una variable dependiente, la cual es "y", esta la tenemos que eliminar, ya que el análisis de componentes principales se realiza con las variables independientes.

Debemos de eliminar esta variable de nuestro conjunto de datos, para ello, corremos el siguiente código:

```
# Corregimos
# Con esto, ya hemos eliminado la variable "y" y nos quedamos con las variables de nuestro in

data2 <- scale(datos[,-16])
view(data2)
```

```r
# Calculamos el determinante de la correlacion
det(cor(data2))
```

```
[1] 0.004667778
```

Al obtener que el determinante es cercano a cero, esto nos indica que los datos son adecuados para realizar un PCA.

```r
# Determinamos los componentes principales
pca_data1 <- princomp(data2)
summary(pca_data1)
```

```
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
```

Una vez realizado los componentes principales, debemos de fijar nuestra atención en el porcentaje acumulado, ya que este es el que nos va a indicar el número de componentes principales que debemos de elegir.

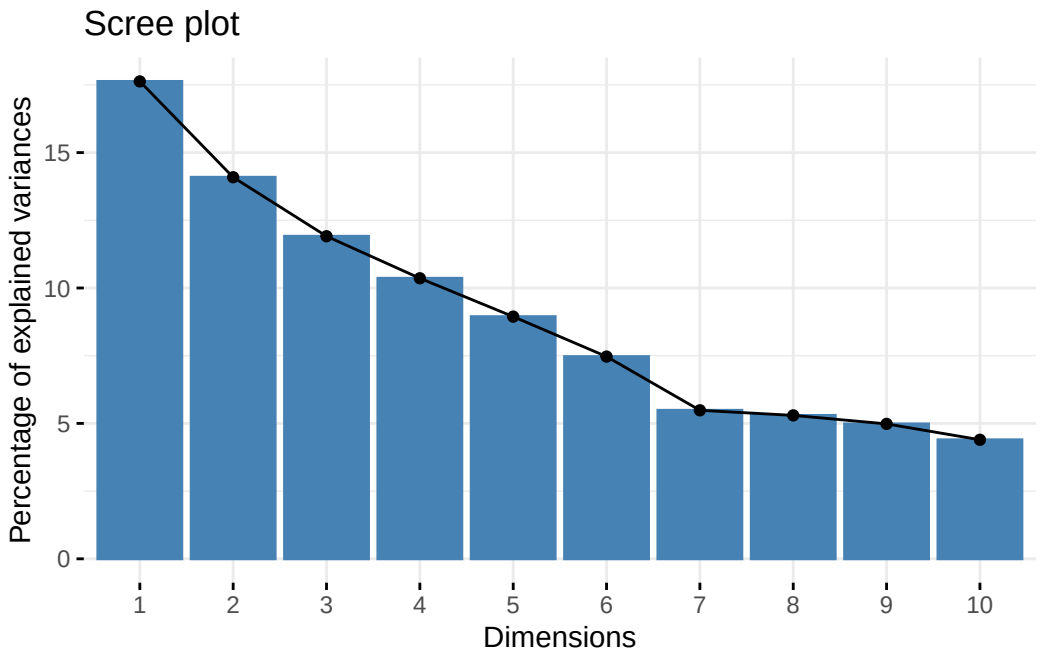
> ⚠ Proporción acumulada
>
> La proporción acumulada es la que nos indica el número de componentes principales que debemos de tomar. Lo ideal es tomar hasta un porcentaje cercano a un 75%-80%. Lo cual nos indica que podemos tomar 6 o 7 componentes.

```r
library(factoextra)
```

Para verificar que hemos elegido el numero correcto de componentes principales, podemos realizar el método del codo, el cual es una manera gráfica de elegir el número de componentes principales.

Gráfica con la varianza

```
# Grafico de sedimentacion
# Para identificar el numero de clusters que debemos de utilizar
fviz_eig(pca_data1, choice="variance")
```

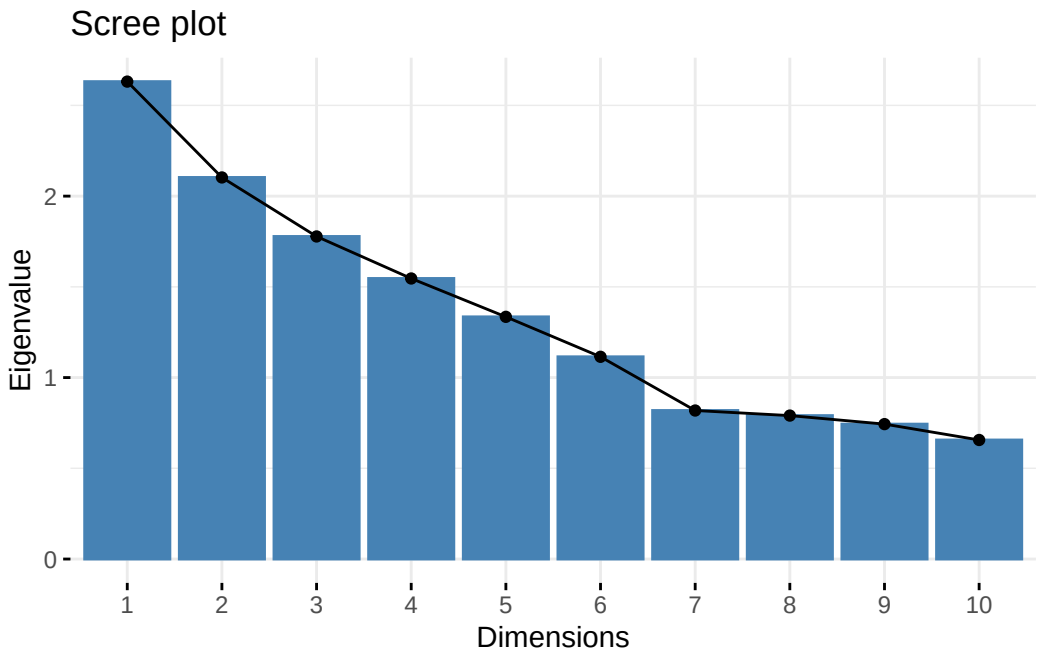



```
# Debemos de ver que, gracias al metodo del codo, debemos de elegir 6 componentes
# principales.
```

La clave está en elegir el componente en donde exista un desnivel notorio.

Gráfica con el eigen-valor

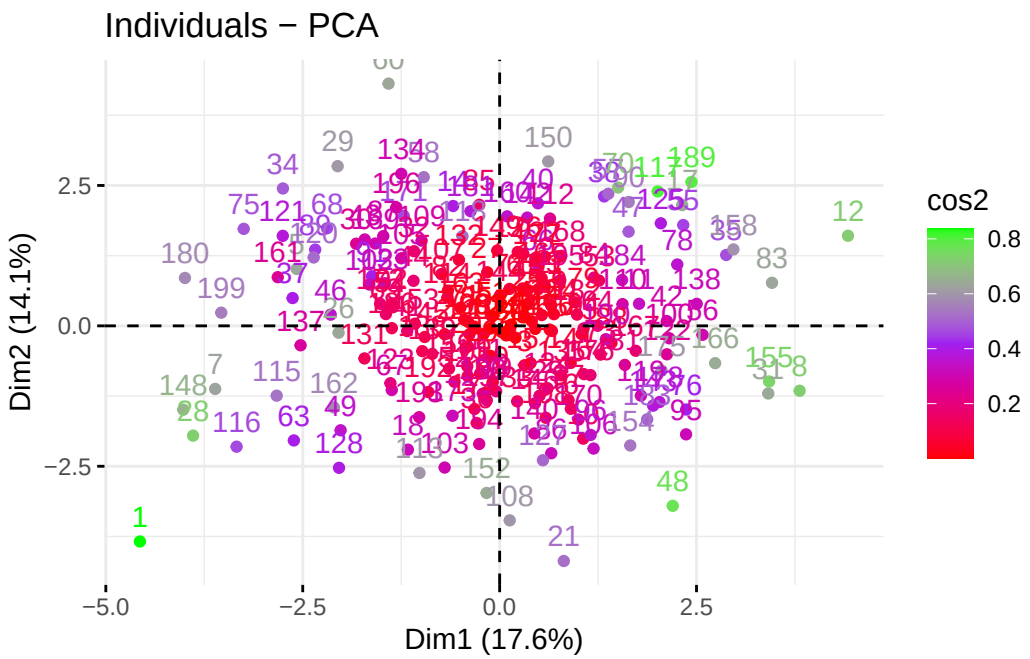
```
# Realizamos la grafica pero con un eigenvalor

# Calculamos los eigenvalores
fviz_eig(pca_data1, choice="eigenvalue")
```




```
# NOTA:
# AL realizar el metodo del codo con el eigen valor, podemos ver que
# tambien los primeros 6 componentes son los que explican la mayor varianza.
# Podemos notar que es lo mismo que cuando graficamos con la varianza.
```

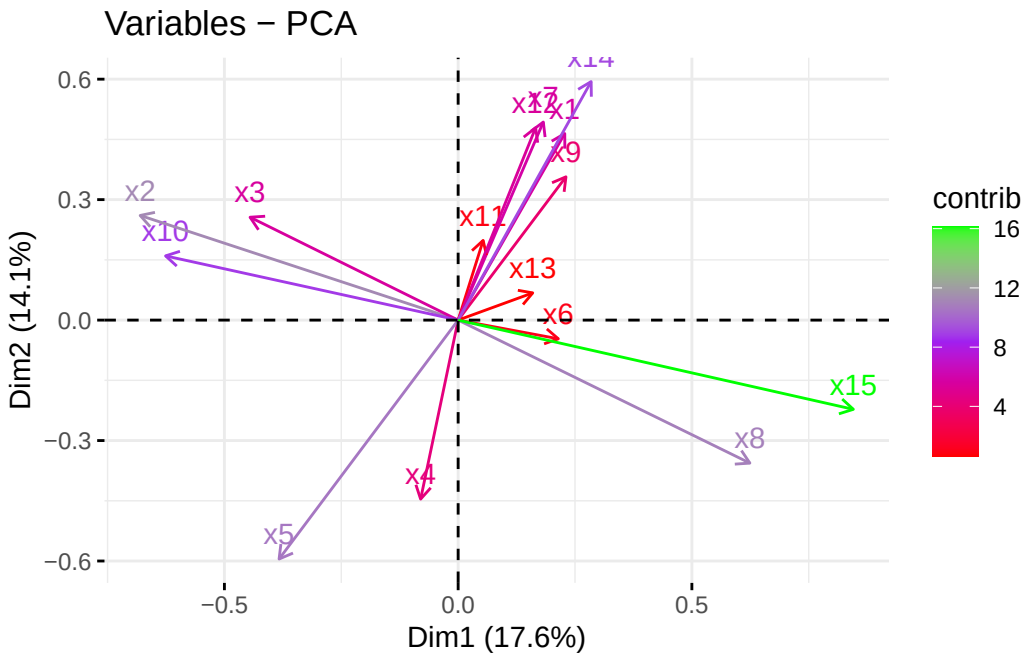
Gráfica de puntuaciones factoriales

```
# Grafico de las puntuaciones factoriales y su representacion
fviz_pca_ind(pca_data1,
             col.ind = "cos2",
             gradient.cols = c("red", "purple", "green"),
             repel = F)
```



> **Interpretación**
>
> Gracias al grafico, podemos ver que esta tecnica no es la óptima para explicar los datos.

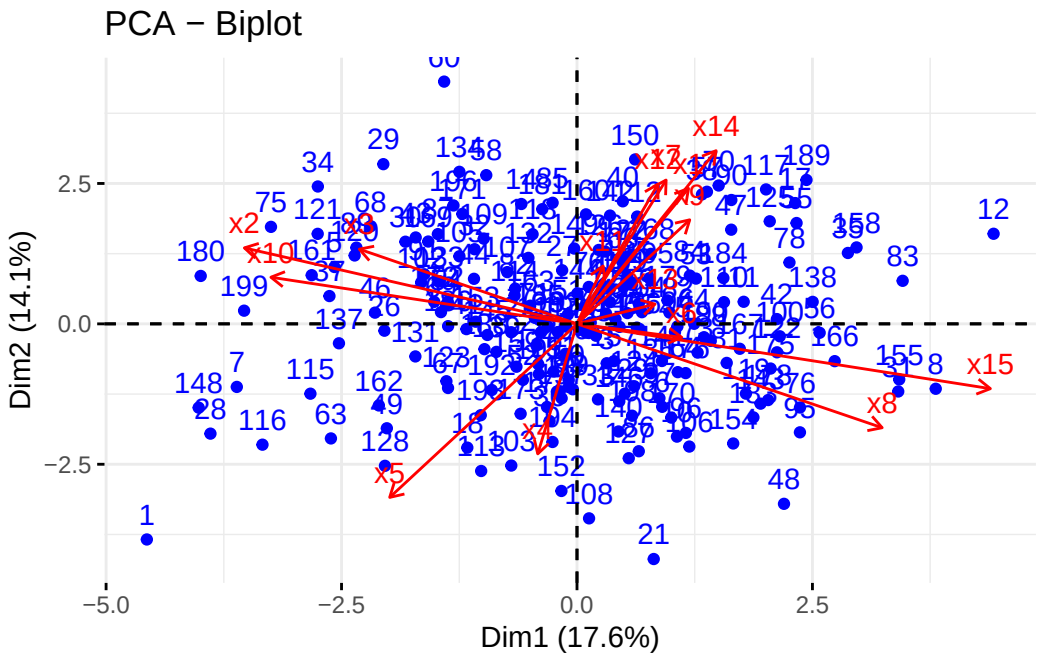
```
# Realizamos otra grafica
fviz_pca_var(pca_data1,
             col.var = "contrib",
             gradient.cols = c("red", "purple", "green"),
             repel = F)
```



Interpretación

Las flechas de color rojo pertenecen a la dimension 2, mientras que las flechas de color verde representan la dimension 1. Las dos contribuyen de manera distinta.

```
fviz_pca_biplot(pca_data1,
                col.var = "red",
                col.ind = "blue")
```



> Combinación de gráficos
>
> Podemos combinar ambos gráficos para visualizar los resultados de una mejor manera.

Pesos de los componentes principales

```
pca2 <- psych::principal(data2, nfactors = 2, residuals = F, rotate = "varimax",
                         scores = T, oblique.scores = F, method = "regression",
                         use = "pairwise", cor = "cor",weight = NULL )
pca2
```

```
Principal Components Analysis
Call: psych::principal(r = data2, nfactors = 2, residuals = F, rotate = "varimax",
    scores = T, oblique.scores = F, method = "regression", use = "pairwise",
    cor = "cor", weight = NULL)
Standardized loadings (pattern matrix) based upon correlation matrix
      RC1   RC2    h2   u2 com
x1   0.04  0.52 0.269 0.73 1.0
x2  -0.73  0.00 0.533 0.47 1.0
x3  -0.51  0.08 0.265 0.73 1.0
```

```
x4   0.09 -0.45 0.206 0.79 1.1
x5  -0.14 -0.70 0.503 0.50 1.1
x6   0.22  0.03 0.048 0.95 1.0
x7  -0.01  0.53 0.278 0.72 1.0
x8   0.71 -0.11 0.519 0.48 1.0
x9   0.09  0.42 0.181 0.82 1.1
x10 -0.64 -0.08 0.419 0.58 1.0
x11 -0.02  0.20 0.042 0.96 1.0
x12 -0.02  0.51 0.258 0.74 1.0
x13  0.12  0.12 0.030 0.97 2.0
x14  0.05  0.66 0.436 0.56 1.0
x15  0.87  0.10 0.768 0.23 1.0

                       RC1  RC2
SS loadings           2.57 2.18
Proportion Var        0.17 0.15
Cumulative Var        0.17 0.32
Proportion Explained  0.54 0.46
Cumulative Proportion 0.54 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  0.14
 with the empirical chi square  770.91  with prob <  1.5e-115

Fit based upon off diagonal values = 0.51
```

```r
# Accedemos a los datos de este data frame
pca2$weights[,1]
```

```
           x1            x2            x3            x4            x5
 0.0008960604 -0.2851378248 -0.2012827708  0.0481355724 -0.0332723838
           x6            x7            x8            x9           x10
 0.0837472648 -0.0202342745  0.2817042861  0.0202814144 -0.2486671202
          x11           x12           x13           x14           x15
-0.0151416330 -0.0236242067  0.0447827540 -0.0013696313  0.3369550463
```

```
# Para el aspecto fisico
pca2$weights[,2]
```

```
          x1           x2           x3           x4           x5           x6
 0.236717095  0.021936576  0.052252404 -0.208155734 -0.315680597  0.008795705
          x7           x8           x9          x10          x11          x12
 0.243092448 -0.071914925  0.189412732 -0.015124597  0.095096723  0.234550137
         x13          x14          x15
 0.051813948  0.301661558  0.017665734
```

```
# Nuevas variables obtenidas, cuya principal caracteristica es que son
# ortogonales, es decir, linealmente independientes.
```

```
# Las variables son las siguientes:
pca2$scores
```

```
               RC1         RC2
  [1,] -1.663181432 -3.47805844
  [2,] -0.363562031 -0.25080118
  [3,]  0.355426863 -0.37948303
  [4,]  0.131447901  0.18641872
  [5,] -1.728595344  0.07626033
  [6,]  0.886144950 -0.74574263
  [7,] -1.791375398 -1.52348776
  [8,]  2.471554555  0.10651225
  [9,] -0.148288881  0.89193562
 [10,]  1.108288252 -1.05150773
 [11,]  0.917048312  0.64596530
 [12,]  2.136851958  2.01301453
 [13,]  0.194033287 -0.81137340
 [14,] -0.869718074  1.23715076
 [15,]  0.044982518  0.47987229
 [16,]  0.084967686  0.01060598
 [17,]  0.793574599  1.89480493
 [18,] -0.119102206 -1.67230798
 [19,] -0.080510505 -0.13905438
 [20,] -0.151593533 -0.34167904
 [21,]  1.509186645 -2.50404963
 [22,]  0.025517491  0.91758816
 [23,]  0.168060496 -0.09080835
 [24,]  0.040314586 -0.06384400
```

```
[25,]  0.105099516  0.32604240
[26,] -1.141803540 -0.53291409
[27,] -0.327694558  0.57688400
[28,] -1.745004932 -2.11925524
[29,] -1.886388920  1.36596967
[30,] -1.408426430  0.53267094
[31,]  2.255198855 -0.01305522
[32,] -0.953839808  0.61028878
[33,]  0.462452562 -0.81267787
[34,] -2.186063854  0.95664069
[35,]  1.335112083  1.44962001
[36,]  0.263479483 -1.17372941
[37,] -1.630151831 -0.26740084
[38,]  0.187192927  1.77358413
[39,]  0.259357062  0.24221556
[40,] -0.264116689  1.50745427
[41,] -0.649503681 -0.32114967
[42,]  1.195849679  0.52694049
[43,] -1.364604928  0.60535110
[44,] -0.826674665  0.27129448
[45,]  0.632862315  0.13811486
[46,] -1.279442465 -0.34897872
[47,]  0.521722448  1.43964201
[48,]  2.057561982 -1.56602701
[49,] -0.694814644 -1.64132533
[50,]  0.198683933 -0.66074941
[51,] -0.096684160  0.08420017
[52,] -0.080792492 -0.76697582
[53,]  0.523728599  0.79783410
[54,] -0.881160378 -0.09244674
[55,]  0.889791101  1.67006207
[56,]  1.516734928  0.47128242
[57,]  0.205675362  1.81559398
[58,] -1.211172109  1.48363719
[59,]  0.285751751  0.37764667
[60,] -1.882018623  2.45253675
[61,]  0.001242478  0.49272459
[62,]  0.229126806 -0.74076759
[63,] -0.990237939 -1.88936199
[64,]  0.712992703  0.27760981
[65,]  0.403242858  0.08942812
[66,]  0.008525364  0.80490821
[67,] -0.500863674 -1.03679961
```

```
 [68,] -1.688732565  0.62609716
 [69,] -0.879917608 -0.18706950
 [70,]  0.249449544  1.91471624
 [71,] -0.283464801 -0.03073237
 [72,] -0.091621925  0.44957214
 [73,]  0.861822200 -0.04294464
 [74,]  0.078035665  0.04721447
 [75,] -2.292492422  0.38474563
 [76,]  1.728119167 -0.42831342
 [77,]  0.094645513 -0.07866099
 [78,]  1.021409038  1.20287383
 [79,]  0.264695360 -0.75680623
 [80,]  0.112260616 -0.07658599
 [81,]  1.084519001 -0.09232373
 [82,] -0.532661091  0.25672190
 [83,]  1.791813516  1.26091207
 [84,]  0.473204015  0.81810229
 [85,] -0.684905270  1.32510017
 [86,]  0.940595870 -1.30853778
 [87,] -1.244494813  0.69632648
 [88,]  0.889744074  0.12173111
 [89,] -1.680528953  0.34984533
 [90,]  0.390870811  1.77882636
 [91,] -1.149286816  0.20564591
 [92,] -0.883995431 -0.05331987
 [93,] -0.056006952 -0.05773772
 [94,]  0.622921653  0.20797725
 [95,]  1.837861719 -0.71191950
 [96,]  1.146132246 -0.98789143
 [97,]  0.831304077 -0.65399469
 [98,] -0.167267711 -0.09652808
 [99,]  0.157844745  0.57471758
[100,]  1.289681991  0.33736814
[101,] -0.009518897 -0.66442164
[102,] -1.131694227  0.10420478
[103,]  0.227820260 -1.77320271
[104,]  0.343636435  0.28669628
[105,] -1.015254442  0.49414549
[106,]  1.227695664 -1.13583579
[107,] -0.653524974  0.42617784
[108,]  0.935352727 -2.19226831
[109,] -0.945359579  0.75278169
[110,]  0.798493248  0.59527668
```

```
[111,] -0.182434415 -0.18049814
[112,] -0.108828086  1.36744887
[113,]  0.068642652 -1.90638751
[114,] -0.505808483  0.17625129
[115,] -1.313357721 -1.42694362
[116,] -1.379364286 -2.12241890
[117,]  0.555141327  1.98190981
[118,] -0.667412652  0.91860021
[119,]  1.337478042 -0.39756339
[120,] -1.656553170  0.25584312
[121,] -1.977257823  0.41552427
[122,]  1.343502445  0.14766768
[123,] -0.540823678 -0.96358806
[124,]  0.625136393 -0.57698845
[125,]  0.718927880  1.62673405
[126,] -0.195155692  0.89523542
[127,]  0.910692513 -1.41201180
[128,] -0.540040515 -2.07462041
[129,]  0.600264251 -0.68426890
[130,]  0.177979990 -0.82184466
[131,] -0.839385964 -0.75379213
[132,] -0.588186259  0.63876238
[133,] -1.024527343  0.16243794
[134,] -1.389595773  1.45835963
[135,]  0.667635654 -0.39040295
[136,] -0.056720053 -0.55261685
[137,] -1.362855606 -0.78420294
[138,]  1.337345287  0.80777575
[139,] -0.366397759 -0.51005677
[140,]  0.729524091 -1.13077035
[141,]  0.681572460 -0.22331639
[142,] -0.278730217  1.31442187
[143,]  1.471062711 -0.47770100
[144,] -0.132713631  0.34929955
[145,] -0.496178664 -0.33851983
[146,]  0.600822286 -0.78493424
[147,]  0.272593112  0.06594810
[148,] -1.931720581 -1.85024125
[149,] -0.350702532  0.85353899
[150,] -0.374593439  2.01477013
[151,]  0.068601982 -0.60043400
[152,]  0.645537289 -1.94493688
[153,] -0.626378015 -0.21433197
```

```
[154,]  1.481035203 -0.99665310
[155,]  2.207929960  0.12640557
[156,]  0.381041005  0.09880233
[157,]  0.390060709 -0.33821528
[158,]  1.363127868  1.53326647
[159,] -0.178529026 -0.63247144
[160,] -0.430415472  1.27189265
[161,] -1.830687715 -0.07070883
[162,] -0.844544670 -1.39959915
[163,] -0.365501252  0.09688615
[164,]  0.827537887 -0.31323932
[165,]  0.094443091  0.66355454
[166,]  1.734208120  0.18287931
[167,]  1.101412087  0.10001821
[168,]  0.139050937  0.97362543
[169,] -1.270278461  0.58877299
[170,]  0.989258008 -0.84731891
[171,] -1.185035259  0.98452974
[172,] -0.966644026 -0.08551493
[173,]  0.055322471 -1.15963929
[174,]  0.185883835 -1.01999761
[175,]  1.376528781 -0.05288940
[176,]  0.879527355 -0.30554160
[177,] -0.025087450  0.86804996
[178,]  1.501287798 -0.41856491
[179,]  0.446654247  0.48227162
[180,] -2.502882395 -0.34160096
[181,] -0.719805613  1.22757385
[182,]  0.513419818  0.34717255
[183,]  1.488350553 -0.65111277
[184,]  0.688111149  0.86961481
[185,] -0.248192785 -0.04273693
[186,] -0.774447598 -0.33076244
[187,]  0.213469367  0.14410373
[188,] -0.451083343 -0.50725241
[189,]  0.759350447  2.18536730
[190,]  0.832397164  0.14776045
[191,]  0.119632381  0.53035188
[192,] -0.225847994 -0.95650810
[193,] -0.180965925 -1.27041253
[194,]  0.374494531 -1.40665842
[195,]  0.238388497  0.69270164
[196,] -1.276731027  1.06037324
```

```
[197,]  0.236446028 -0.88610798
[198,]  0.742677578 -0.91861134
[199,] -2.084771954 -0.63576583
[200,]  0.063069234 -0.36840522
```

Segundo documento: Población USA

> **!** Análisis de Componentes Principales
>
> Se realizará un análisis de componentes principales para la población en los Estados Unidos. Se realizará un análisis para el año 2000 y otro para el año 2001.

```
# Leemos los datos
require(tidyverse)
require(readxl)
```

```
Cargando paquete requerido: readxl
```

```
require(factoextra)
```

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

summary(datos_generales)
```

```
   State           Census Resident Total Population - AB:Qr-1-2000
 Length:51          Min.   :  493782
 Class :character   1st Qu.: 1502608
 Mode  :character   Median : 4012012
                    Mean   : 5518077
                    3rd Qu.: 6214791
                    Max.   :33871648
 Resident Total Population Estimate - Jul-1-2000
 Min.   :  494001
 1st Qu.: 1505918
 Median : 4023438
```

```
Mean   : 5531856
3rd Qu.: 6223511
Max.   :34000446
Resident Total Population Estimate - Jul-1-2001
Min.   :  494423
1st Qu.: 1517120
Median : 4063011
Mean   : 5584253
3rd Qu.: 6247024
Max.   :34501130
Net Domestic Migration - Jul-1-2000 Net Domestic Migration - Jul-1-2001
Min.   :-44761                      Min.   :-204875
1st Qu.: -2264                      1st Qu.: -11716
Median :  -282                      Median :  -1568
Mean   :     0                      Mean   :      0
3rd Qu.:  1794                      3rd Qu.:  10176
Max.   : 31461                      Max.   : 205303
Federal/Civilian Movement from Abroad - Jul-1-2000
Min.   :  0.00
1st Qu.:  5.00
Median : 12.00
Mean   : 38.94
3rd Qu.: 41.00
Max.   :336.00
Federal/Civilian Movement from Abroad - Jul-1-2001
Min.   :-1515.0
1st Qu.: -208.5
Median :  -66.0
Mean   : -199.9
3rd Qu.:  -28.5
Max.   :   -1.0
Net International Migration - Jul-1-2000
Min.   :  100
1st Qu.:  706
Median : 1571
Mean   : 5389
3rd Qu.: 4557
Max.   :71852
Net International Migration - Jul-1-2001 Period Births - Jul-1-2000
Min.   :   368                           Min.   :  1480
1st Qu.:  2956                           1st Qu.:  4905
Median :  6090                           Median : 13360
Mean   : 20882                           Mean   : 19405
```

```
3rd Qu.: 18823                          3rd Qu.: 20768
Max.   :271841                          Max.   :129777
Period Births - Jul-1-2001 Period Deaths - Jul-1-2000
Min.   :  6130             Min.   :  679
1st Qu.: 20125             1st Qu.: 2908
Median : 55402             Median : 8118
Mean   : 79467             Mean   :11053
3rd Qu.: 85186             3rd Qu.:13136
Max.   :530349             Max.   :54040
Period Deaths - Jul-1-2001 Resident Under 65 Population Estimate - Jul-1-2000
Min.   :  2949             Min.   :  436178
1st Qu.: 12562             1st Qu.: 1316530
Median : 35149             Median : 3535770
Mean   : 47752             Mean   : 4844195
3rd Qu.: 56776             3rd Qu.: 5416612
Max.   :231693             Max.   :30389907
Resident Under 65 Population Estimate - Jul-1-2001
Min.   :  436217
1st Qu.: 1326894
Median : 3559447
Mean   : 4892273
3rd Qu.: 5440513
Max.   :30845002
Resident 65 Plus Population Estimate - Jul-1-2000
Min.   :  35957
1st Qu.: 187602
Median : 470475
Mean   : 687660
3rd Qu.: 772224
Max.   :3610539
Resident 65 Plus Population Estimate - Jul-1-2001 Residual - Jul-1-2000
Min.   :  36856                                   Min.   :-678.0
1st Qu.: 190297                                   1st Qu.: -88.5
Median : 470472                                   Median : -28.0
Mean   : 691979                                   Mean   :   0.0
3rd Qu.: 778044                                   3rd Qu.:  45.5
Max.   :3656128                                   Max.   :1043.0
Residual - Jul-1-2001
Min.   :-2175.0
1st Qu.: -368.0
Median : -140.0
Mean   :    0.0
3rd Qu.:   46.5
```

```
 Max.   : 3348.0
```

```
view(datos_generales)
```

Año 2000

Importamos nuestra base de datos con las columnas que correspondan al año 2000, incluyendo el Estado, ya que este será nuestra variable dependiente.

```
# Base de datos 2000
data_2000 <- datos_generales[,c(1,2,3,5,7,9,11,13,15,17,19)]
```

> **!** Normalizar los datos
>
> Recordemos que para poder realizar este análisis de componentes principales, es importante normalizar los datos.

```
# Normalizamos los datos
norm_2000 <- scale(data_2000[,-1])

det(cor(norm_2000))
```

```
[1] -8.80948e-41
```

Determinante que tiende a cero, lo cual nos indica que los datos son optimos para un analisis de componentes principales.

```
pca_norm_2000 <- princomp(norm_2000)

summary(pca_norm_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
```

Debemos de quedarnos con 2 componentes, ya que hasta el componente 2 explica el 87.58% de la varianza. Esto ya es una condición suficiente para poder elegir los componentes con los que nos vamos a quedar.

Calculamos el factor de adecuacion muestral de Kaiser.

```
psych::KMO(norm_2000)
```

```
Error in solve.default(r) :
  sistema es computacionalmente singular: número de condición recíproco = 3.40445e-18
```

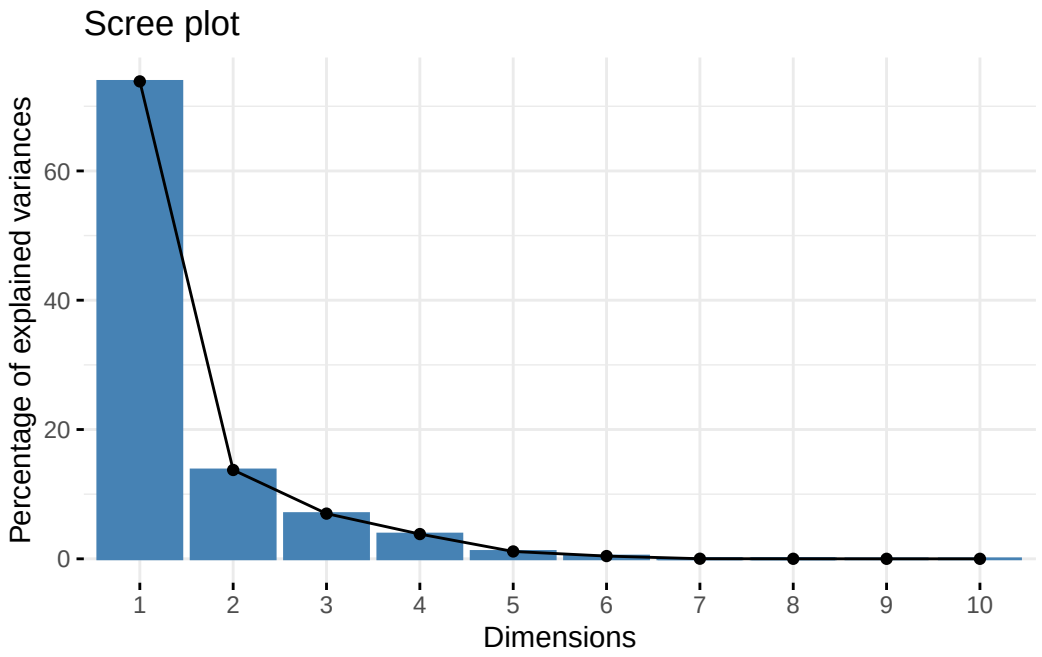
```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = norm_2000)
Overall MSA =  0.5
MSA for each item =
   Census Resident Total Population - AB:Qr-1-2000
                                                0.5
   Resident Total Population Estimate - Jul-1-2000
                                                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
```

El valor obtenido al correr KMO es de 0.5, no es el valor optimo, ya que, el valor optimo seria de 0.6 o mayor, pero es un valor que nos resulta útil.

Para verificar que hemos elegido el numero correcto de componentes principales, podemos realizar el método del codo, el cual es una manera gráfica de elegir el número de componentes principales.

Gráfica con la varianza

```
# Grafico de sedimentacion
# Para identificar el numero de clusters que debemos de utilizar
fviz_eig(pca_norm_2000, choice="variance")
```

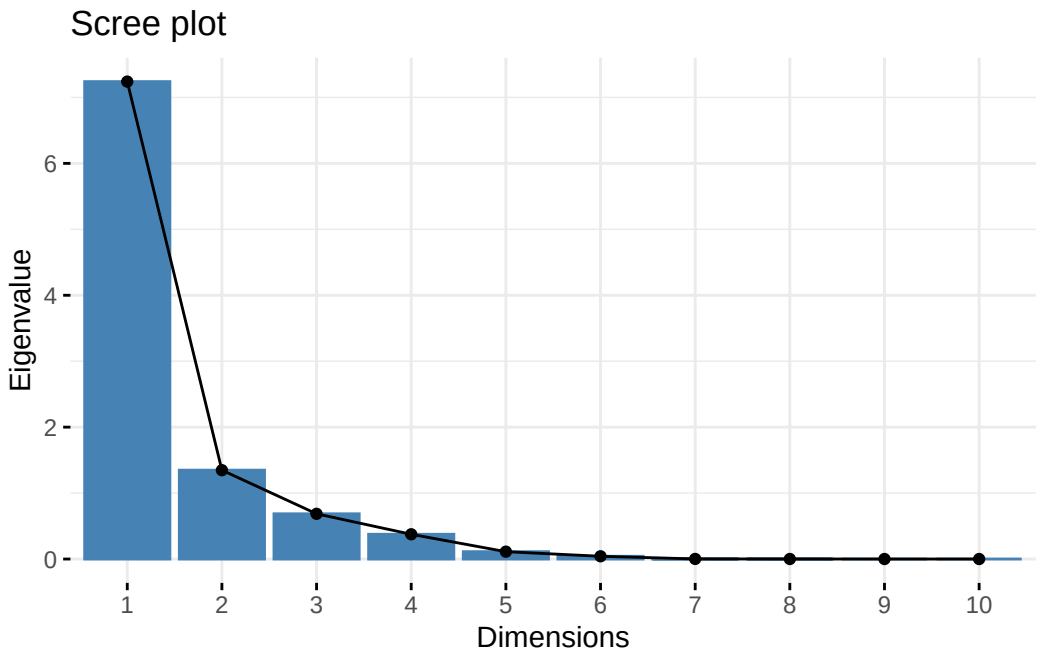



```
# Debemos de ver que, gracias al metodo del codo, debemos de elegir 2 componentes
# principales.
```

La clave está en elegir el componente en donde exista un desnivel notorio o si notamos un cambio brusco de un componente a otro.

Gráfica con el eigen-valor

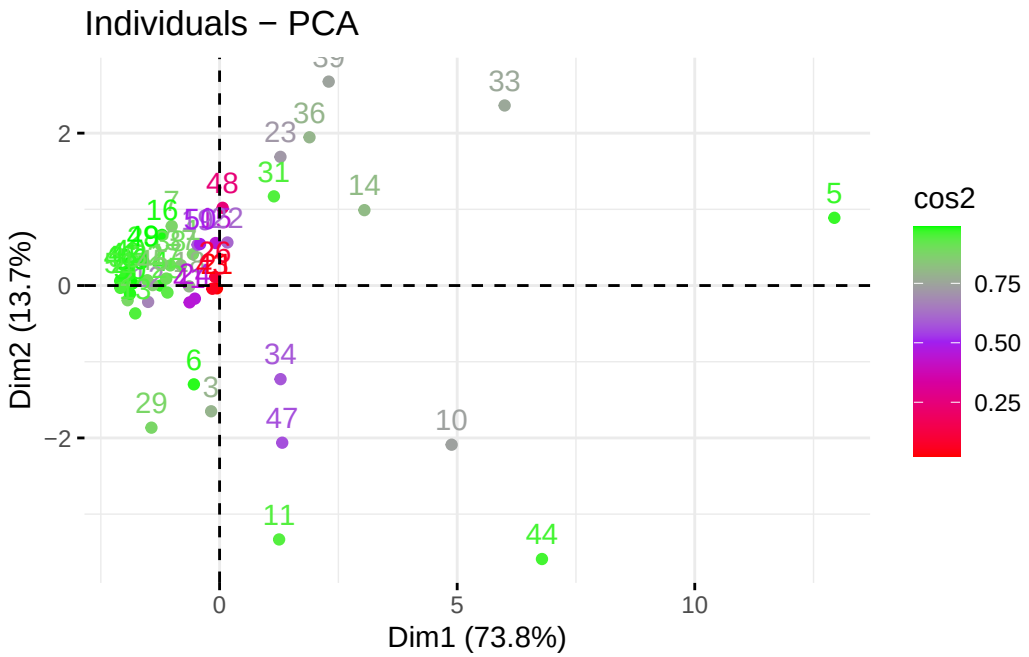
```
# Realizamos la grafica pero con un eigenvalor

# Calculamos los eigenvalores
fviz_eig(pca_norm_2000, choice="eigenvalue")
```




```
# NOTA:
# AL realizar el metodo del codo con el eigen valor, podemos ver que
# tambien los primeros 2 componentes son los que explican la mayor varianza.
# Podemos notar que es lo mismo que cuando graficamos con la varianza.
```

Gráfica de puntuaciones factoriales

```
# Grafico de las puntuaciones factoriales y su representacion
fviz_pca_ind(pca_norm_2000,
             col.ind = "cos2",
             gradient.cols = c("red", "purple", "green"),
             repel = F)
```

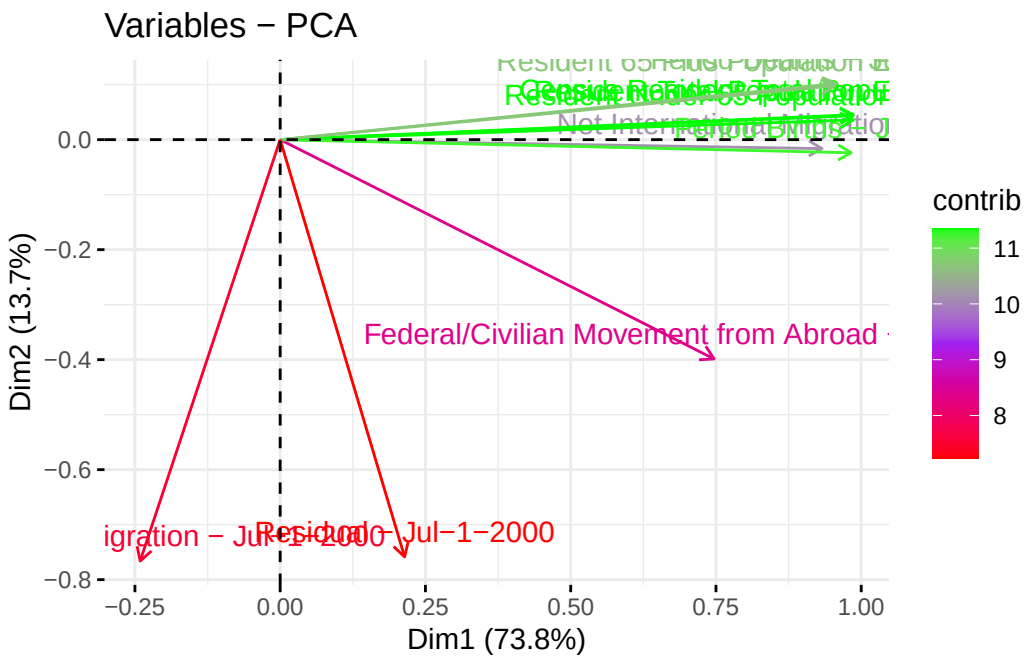


Interpretación

Gracias al gráfico, podemos apreciar que, en este caso, se han representado de manera correcta los componentes. Vemos que aquellos puntos que tienden a ser de color rojo son aquellas observaciones que no han sido representadas de la mejor manera, mientras que los puntos verdes son aquellos que han sido representados de mejor manera.

Contribuciones

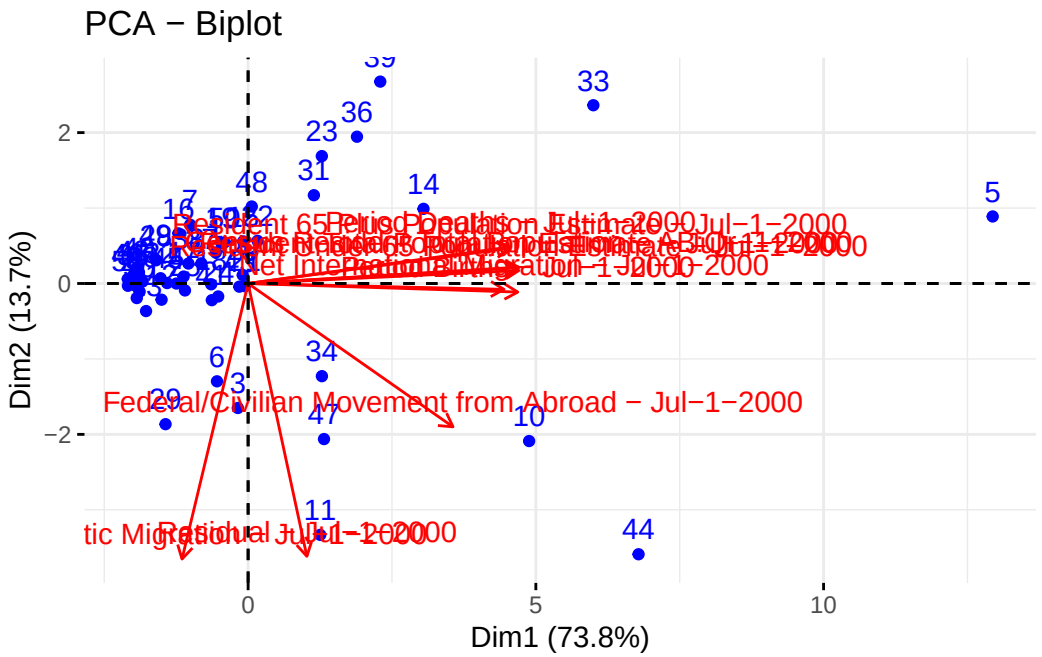
```
# Realizamos otra grafica
fviz_pca_var(pca_norm_2000,
             col.var = "contrib",
             gradient.cols = c("red", "purple", "green"),
             repel = F)
```



Interpretación

Las flechas de color rojo pertenecen a la dimension 2, mientras que las flechas de color verde representan la dimension 1. Las dos contribuyen de manera distinta.

```
fviz_pca_biplot(pca_norm_2000,
                col.var = "red",
                col.ind = "blue")
```



> Combinación de gráficos
>
> Podemos combinar ambos gráficos para visualizar los resultados de una mejor manera.

Pesos de los componentes principales

```
pca_2000 <- psych::principal(norm_2000, nfactors = 2, residuals = F, rotate = "varimax",
                             scores = T, oblique.scores = F, method = "regression",
                             use = "pairwise", cor = "cor",weight = NULL )
pca_2000
```

```
Principal Components Analysis
Call: psych::principal(r = norm_2000, nfactors = 2, residuals = F,
    rotate = "varimax", scores = T, oblique.scores = F, 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
```

```r
# Accedemos a los datos de este data frame
pca_2000$weights[,1]
```

```
   Census Resident Total Population - AB:Qr-1-2000
                                          0.9968697
   Resident Total Population Estimate - Jul-1-2000
                                          0.9969223
              Net Domestic Migration - Jul-1-2000
                                         -0.2636335
Federal/Civilian Movement from Abroad - Jul-1-2000
                                          0.7433061
          Net International Migration - Jul-1-2000
                                          0.9410233
                       Period Births - Jul-1-2000
                                          0.9916210
                       Period Deaths - Jul-1-2000
                                          0.9683729
Resident Under 65 Population Estimate - Jul-1-2000
                                          0.9968932
```

```
Resident 65 Plus Population Estimate - Jul-1-2000
                                          0.9655511
                              Residual - Jul-1-2000
                                          0.1960313
```

```
pca_2000$weights[,2]
```

```
   Census Resident Total Population - AB:Qr-1-2000
                                        -0.01959089
   Resident Total Population Estimate - Jul-1-2000
                                        -0.01779679
               Net Domestic Migration - Jul-1-2000
                                         0.76709127
Federal/Civilian Movement from Abroad - Jul-1-2000
                                         0.42229304
          Net International Migration - Jul-1-2000
                                         0.04124540
                        Period Births - Jul-1-2000
                                         0.05018618
                        Period Deaths - Jul-1-2000
                                        -0.07748049
Resident Under 65 Population Estimate - Jul-1-2000
                                        -0.01011621
 Resident 65 Plus Population Estimate - Jul-1-2000
                                        -0.07357619
                             Residual - Jul-1-2000
                                         0.77152787
```

```
# Las variables son las siguientes:
pca_2000$scores
```

```
             RC1          RC2
 [1,] -1.5127071 -0.521165011
 [2,] -5.2605020  0.088078420
 [3,] -0.5341723  1.921398070
 [4,] -3.3855968 -0.088306450
 [5,] 35.1759366 -0.116369481
 [6,] -1.5094060  1.481098884
 [7,] -2.7135464 -0.985708499
 [8,] -5.4131588 -0.138083070
 [9,] -5.3739938 -0.159813920
[10,] 13.2006094  2.796752770
[11,]  3.3010060  3.994742052
[12,] -4.0920322  0.144177579
[13,] -4.8297127  0.301765249
[14,]  8.3101230 -0.941813590
```

```
[15,] -0.2132614 -0.659945915
[16,] -3.2703445 -0.865380352
[17,] -3.0424731 -0.187770338
[18,] -1.7423632 -0.033263497
[19,] -1.2442065 -0.659927945
[20,] -4.9781408 -0.162432446
[21,] -0.1524196  0.038787513
[22,]  0.4641848 -0.650492347
[23,]  3.5317840 -1.887370946
[24,] -1.4179092  0.164403079
[25,] -2.9826801  0.029385425
[26,] -0.2502966 -0.137183567
[27,] -5.2833345 -0.097555353
[28,] -4.3365309 -0.474604471
[29,] -3.9546473  2.083128252
[30,] -5.1253077 -0.008284433
[31,]  3.1414489 -1.289918307
[32,] -4.1298008 -0.190583013
[33,] 16.3700351 -2.340194651
[34,]  3.4464967  1.531472479
[35,] -5.4104720 -0.196846202
[36,]  5.1997056 -2.144799530
[37,] -2.1972938 -0.370907570
[38,] -2.8093637 -0.383989723
[39,]  6.3210727 -2.972128157
[40,] -5.1477668 -0.308591011
[41,] -1.7197493  0.213935136
[42,] -5.4218284 -0.268625283
[43,] -0.4155914  0.038433890
[44,] 18.3208364  4.691183009
[45,] -3.8227564 -0.107390915
[46,] -5.6807144 -0.219707513
[47,]  3.5187088  2.511416312
[48,]  0.2049286 -1.191068125
[49,] -4.3590920 -0.489490073
[50,] -1.0998239 -0.667074833
[51,] -5.6738804 -0.113371581
```

Año 2001

Importamos nuestra base de datos con las columnas que correspondan al año 2001, incluyendo el Estado, ya que este será nuestra variable dependiente.

```
# Base de datos 2001
data_2001 <- datos_generales[,c(1,4,6,8,10,12,14,16,18,20)]
view(data_2001)
```

```
# Normalizamos los datos
norm_2001 <- scale(data_2001[,-1])
view(norm_2001)

det(cor(norm_2001))
```

```
[1] -4.747386e-25
```

El determinante que tiende a cero, lo cual nos indica que nuestros datos son optimos para un análisis de componentes principales.

```
pca_norm_2001 <- princomp(norm_2001)

summary(pca_norm_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
```

Debemos de quedarnos con 2 componentes, ya que hasta el componente 2 explica el 89.84% de la varianza. Esto ya es una condición suficiente para poder elegir los componentes con los que nos vamos a quedar.

Calculamos el factor de adecuacion muestral de Kaiser.

```r
psych::KMO(norm_2001)
```

```
Error in solve.default(r) :
  sistema es computacionalmente singular: número de condición recíproco = 4.27828e-18
```

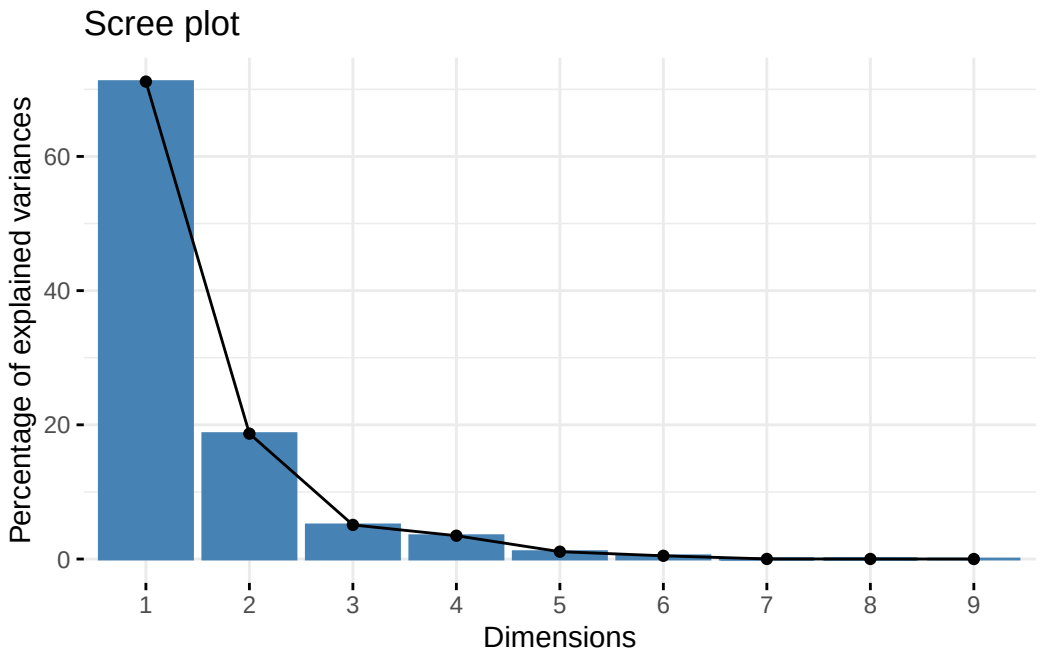
```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = norm_2001)
Overall MSA =  0.5
MSA for each item =
   Resident Total Population Estimate - Jul-1-2001
                                                0.5
               Net Domestic Migration - Jul-1-2001
                                                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
```

El valor obtenido al correr KMO es de 0.5, no es el valor optimo, ya que, el valor optimo seria de 0.6 o mayor, pero es un valor que nos resulta útil.

Para verificar que hemos elegido el numero correcto de componentes principales, podemos realizar el método del codo, el cual es una manera gráfica de elegir el número de componentes principales.

Gráfica con la varianza

```r
# Grafico de sedimentacion
# Para identificar el numero de clusters que debemos de utilizar
fviz_eig(pca_norm_2001, choice="variance")
```

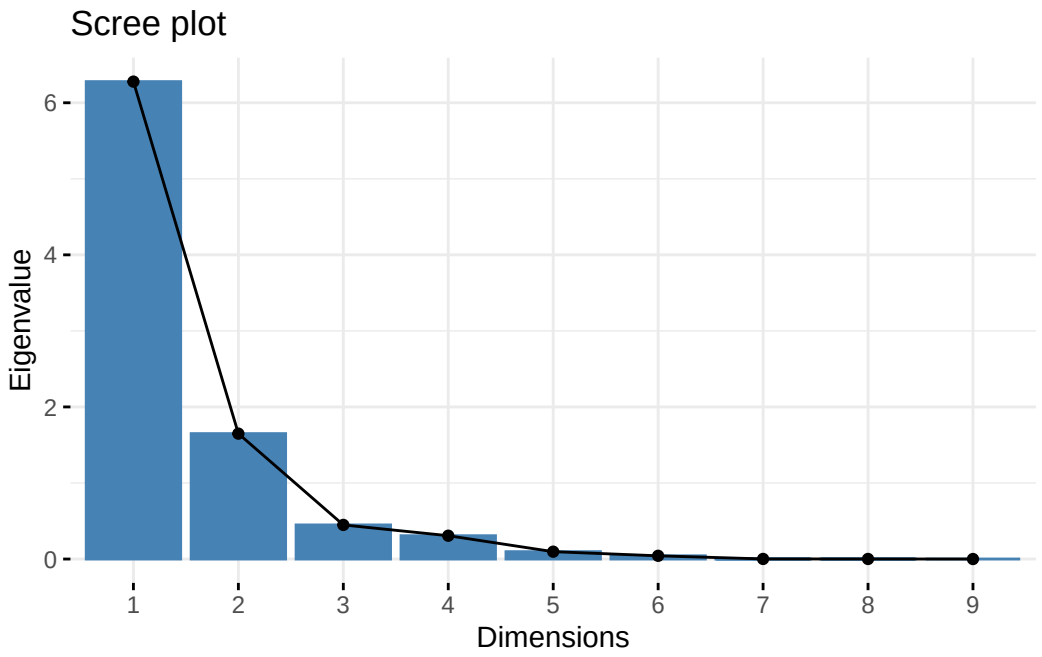



```
# Debemos de ver que, gracias al metodo del codo, debemos de elegir 2 componentes
# principales.
```

La clave está en elegir el componente en donde exista un desnivel notorio o si notamos un cambio brusco de un componente a otro.

Gráfica con el eigen-valor

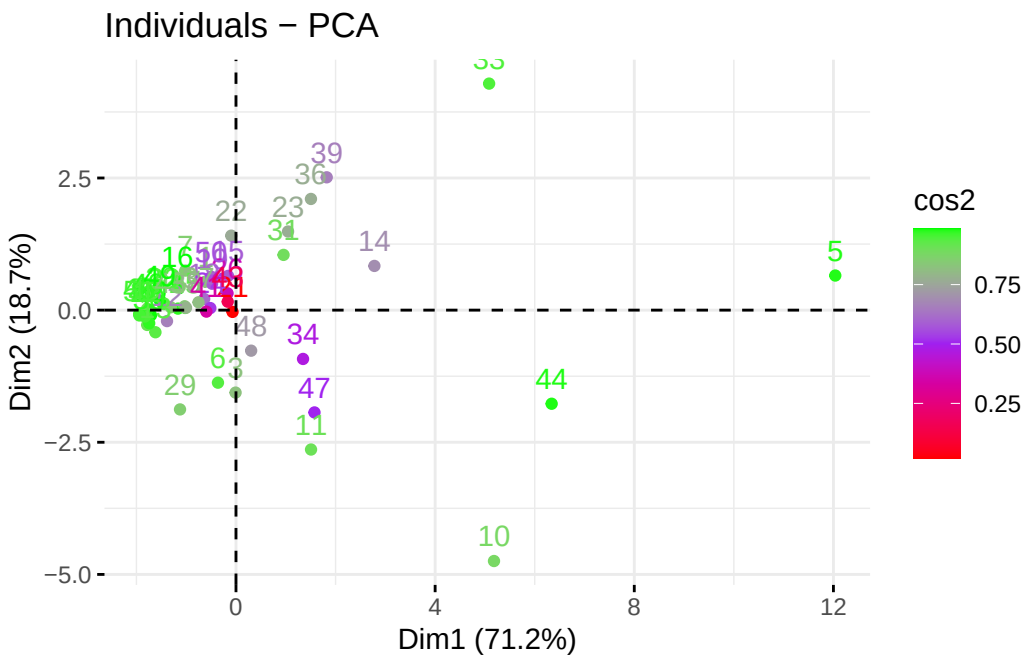
```
# Realizamos la grafica pero con un eigenvalor

# Calculamos los eigenvalores
fviz_eig(pca_norm_2001, choice="eigenvalue")
```




```
# NOTA:
# AL realizar el metodo del codo con el eigen valor, podemos ver que
# tambien los primeros 2 componentes son los que explican la mayor varianza.
# Podemos notar que es lo mismo que cuando graficamos con la varianza.
```

Gráfica de puntuaciones factoriales

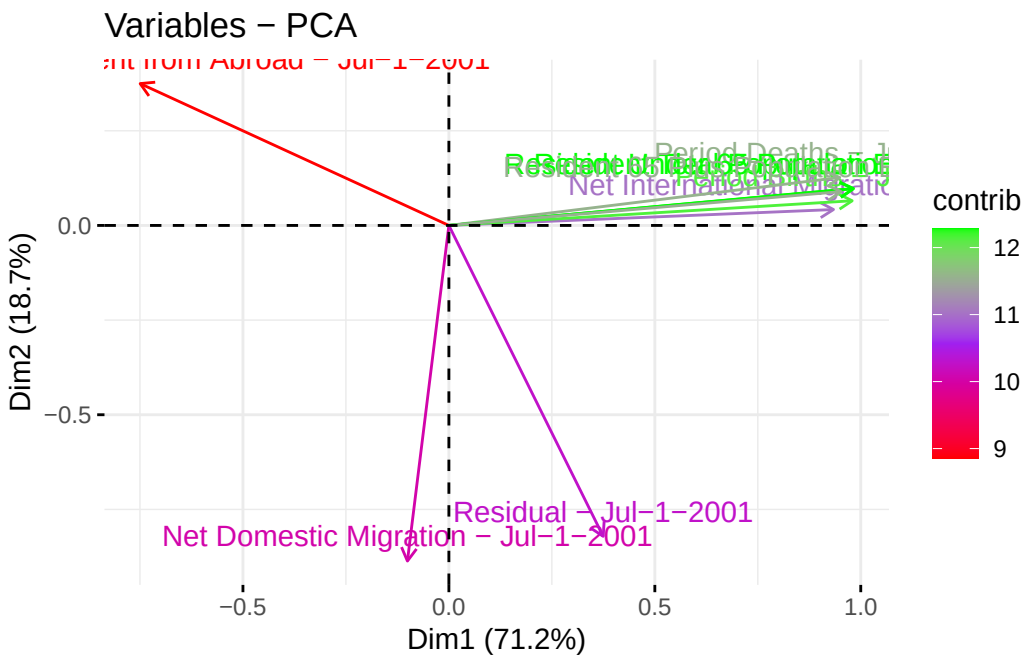
```
# Grafico de las puntuaciones factoriales y su representacion
fviz_pca_ind(pca_norm_2001,
             col.ind = "cos2",
             gradient.cols = c("red", "purple", "green"),
             repel = F)
```



> **Interpretación**
>
> Gracias al gráfico, podemos apreciar que, en este caso, se han representado de manera correcta los componentes. Vemos que aquellos puntos que tienden a ser de color rojo son aquellas observaciones que no han sido representadas de la mejor manera, mientras que los puntos verdes son aquellos que han sido representados de mejor manera.

Contribuciones

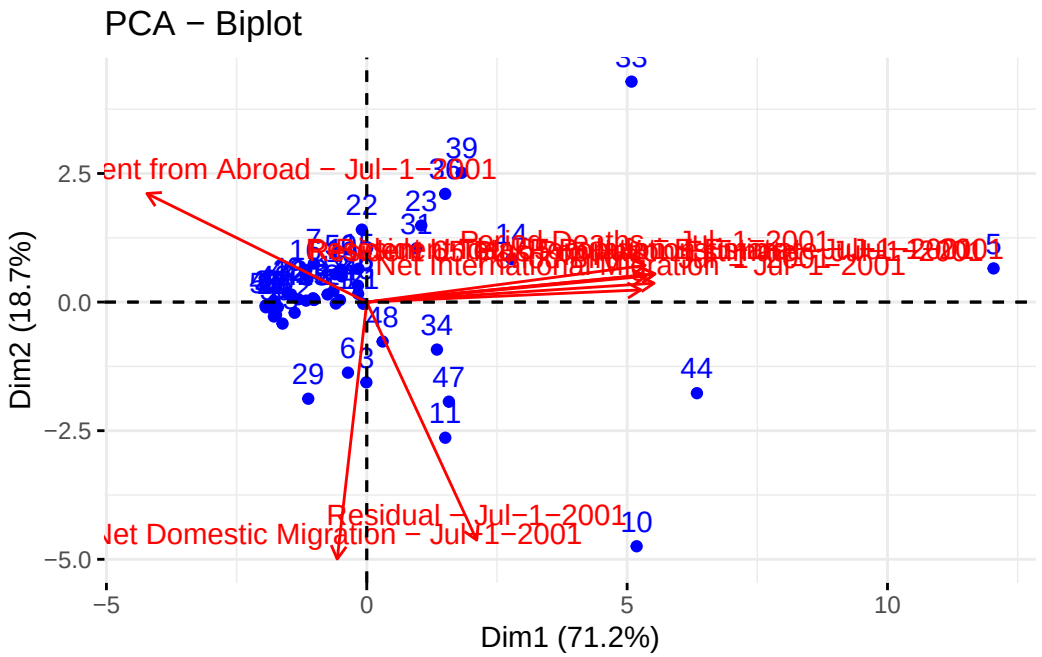
```
# Realizamos otra grafica
fviz_pca_var(pca_norm_2001,
             col.var = "contrib",
             gradient.cols = c("red", "purple", "green"),
             repel = F)
```



Interpretación

Las flechas de color rojo pertenecen a la dimension 2, mientras que las flechas de color verde representan la dimension 1. Las dos contribuyen de manera distinta.

```
fviz_pca_biplot(pca_norm_2001,
                col.var = "red",
                col.ind = "blue")
```



> **Combinación de gráficos**
>
> Podemos combinar ambos gráficos para visualizar los resultados de una mejor manera.

Pesos de los componentes principales

```
pca_2001 <- psych::principal(norm_2001, nfactors = 2, residuals = F, rotate = "varimax",
                             scores = T, oblique.scores = F, method = "regression",
                             use = "pairwise", cor = "cor",weight = NULL )
pca_2001
```

```
Principal Components Analysis
Call: psych::principal(r = norm_2001, nfactors = 2, residuals = F,
    rotate = "varimax", scores = T, oblique.scores = F, 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
```

```r
# Accedemos a los datos de este data frame
pca_2001$weights[,1]
```

```
   Resident Total Population Estimate - Jul-1-2001
                                          0.9960902
              Net Domestic Migration - Jul-1-2001
                                         -0.2169152
Federal/Civilian Movement from Abroad - Jul-1-2001
                                         -0.7015148
          Net International Migration - Jul-1-2001
                                          0.9406167
                       Period Births - Jul-1-2001
                                          0.9890650
                       Period Deaths - Jul-1-2001
                                          0.9669080
Resident Under 65 Population Estimate - Jul-1-2001
                                          0.9961285
 Resident 65 Plus Population Estimate - Jul-1-2001
                                          0.9649681
                             Residual - Jul-1-2001
```

```
0.2676205
```

```
pca_2001$weights[,2]
```

```
   Resident Total Population Estimate - Jul-1-2001
                                         0.031603481
                Net Domestic Migration - Jul-1-2001
                                         0.874405250
Federal/Civilian Movement from Abroad - Jul-1-2001
                                        -0.473297982
           Net International Migration - Jul-1-2001
                                         0.080291348
                          Period Births - Jul-1-2001
                                         0.063045013
                          Period Deaths - Jul-1-2001
                                        -0.002562161
Resident Under 65 Population Estimate - Jul-1-2001
                                         0.030980891
 Resident 65 Plus Population Estimate - Jul-1-2001
                                         0.035217378
                               Residual - Jul-1-2001
                                         0.871359634
```

```
# Las variables son las siguientes:
pca_2001$scores
```

```
                RC1          RC2
 [1,] -1.4291841341 -0.895589927
 [2,] -4.5218019005 -0.225123863
 [3,] -0.2824457258  2.003493806
 [4,] -2.9196880400 -0.419243592
 [5,] 30.3149688256  3.109167975
 [6,] -1.1377761088  1.646116160
 [7,] -2.4271045796 -1.296376113
 [8,] -4.6451810435 -0.505020445
 [9,] -4.6828920563 -0.539294950
[10,] 12.2059011851  7.805516259
[11,]  3.3391111202  3.886120105
[12,] -3.5103714637 -0.187796367
[13,] -4.1291024130  0.007248103
[14,]  7.1172472823 -0.163081667
[15,] -0.2891830258 -0.885518029
[16,] -2.8669196336 -1.076939998
```

```
[17,] -2.5735488746 -0.433211616
[18,] -1.5666950230 -0.482587488
[19,] -1.1198693483 -0.798888935
[20,] -4.2999265633 -0.435013511
[21,] -0.1783808397  0.020322413
[22,] -0.0005745319 -1.842910534
[23,]  2.8756393843 -1.569909740
[24,] -1.2857362018 -0.219614930
[25,] -2.5535394289 -0.388290829
[26,] -0.3704120562 -0.465809778
[27,] -4.5271735883 -0.432530571
[28,] -3.7187359157 -0.679554291
[29,] -3.1357051016  2.047016827
[30,] -4.4153947011 -0.252894045
[31,]  2.5732276573 -1.030518910
[32,] -3.6091967232 -0.620880183
[33,] 13.4790034088 -3.842743194
[34,]  3.2249490627  1.629815253
[35,] -4.6399425997 -0.507231660
[36,]  4.1329412793 -2.211222124
[37,] -1.8633577761 -0.438310335
[38,] -2.5094843875 -0.389091106
[39,]  4.9957903006 -2.633449571
[40,] -4.4430947999 -0.606538265
[41,] -1.4946138247 -0.159111905
[42,] -4.6090754460 -0.444260502
[43,] -0.3884583164 -0.260214889
[44,] 15.6117874789  4.358583262
[45,] -3.2776626554 -0.499760309
[46,] -4.8833313888 -0.550632272
[47,]  3.6316197358  3.006730211
[48,]  0.6355026177  1.086291021
[49,] -3.8038639562 -0.733952684
[50,] -1.1452304803 -0.976110868
[51,] -4.8830346852 -0.507191398
```