Felicidad en el mundo. Análisis de componentes principales

Julio David Ruiz Mendoza

lunes 07 de junio de 2021

Índice

1 Introducción

En este documento vamos a tratar un conjunto datos sobre la felicidad en distintos paises alrededor del planeta. Vamos a importar el archivo y preparar el dataframe seleccionando los datos que nos sean útiles, después vamos a hacer una introducción a los datos y por último definir brevemente qué es un PCA.

He elegido este tema porque parece interesante saber los factores que influyen en la felicidad y en qué países la población es más feliz.

2 Lectura y preparación de datos

Trabajamos con el fichero: `world-happiness-report-2021.csv`
Proveniente de la base de datos Kaggle
Nº de observaciones: 149
Nº de variables: 20

Vista de datos:

Pais	Region	Felicidad	PIB	Apoyo.social	Esp.vida	Libertad	Generosidad	Corrupcion
Finland	EUW	7.842	10.775	0.954	72.000	0.949	-0.098	0.186
Denmark	EUW	7.620	10.933	0.954	72.700	0.946	0.030	0.179
Switzerland	EUW	7.571	11.117	0.942	74.400	0.919	0.025	0.292
Iceland	EUW	7.554	10.878	0.983	73.000	0.955	0.160	0.673
Netherlands	EUW	7.464	10.932	0.942	72.400	0.913	0.175	0.338
Norway	EUW	7.392	11.053	0.954	73.300	0.960	0.093	0.270
Sweden	EUW	7.363	10.867	0.934	72.700	0.945	0.086	0.237
Luxembourg	EUW	7.324	11.647	0.908	72.600	0.907	-0.034	0.386
New Zealand	AMN	7.277	10.643	0.948	73.400	0.929	0.134	0.242
Austria	EUW	7.268	10.906	0.934	73.300	0.908	0.042	0.481
Australia	AMN	7.183	10.796	0.940	73.900	0.914	0.159	0.442
Israel	AFN	7.157	10.575	0.939	73.503	0.800	0.031	0.753

3 Análisis de los datos

3.1 Resumen estadístico

Vamos a realizar un breve resumen de las variables numéricas que tenemos:

Tabla 3: Resumen estadístico de las variables

	Min	Q1	Mediana	Media	Q3	Max	Var
Felicidad	2.52	4.85	5.53	5.53	6.26	7.84	1.15
PIB	6.64	8.54	9.57	9.43	10.42	11.65	1.34
Apoyo.social	0.46	0.75	0.83	0.81	0.90	0.98	0.01
Esp.vida	48.48	59.80	66.60	64.99	69.60	76.95	45.73
Libertad	0.38	0.72	0.80	0.79	0.88	0.97	0.01
Generosidad	-0.29	-0.13	-0.04	-0.02	0.08	0.54	0.02
Corrupcion	0.08	0.67	0.78	0.73	0.84	0.94	0.03

3.2 Vista de los datos

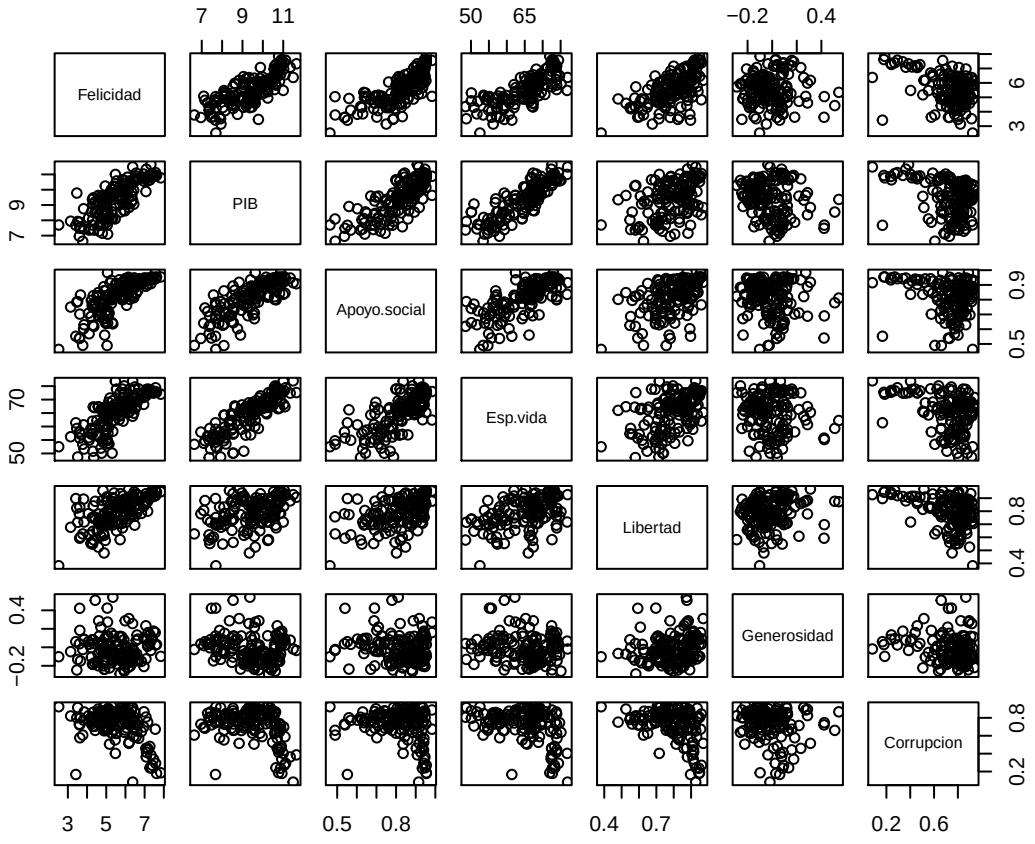

Figura 1: Representación de todas las variables.

No podemos ignorar que hay variables que están relacionadas de algun modo .

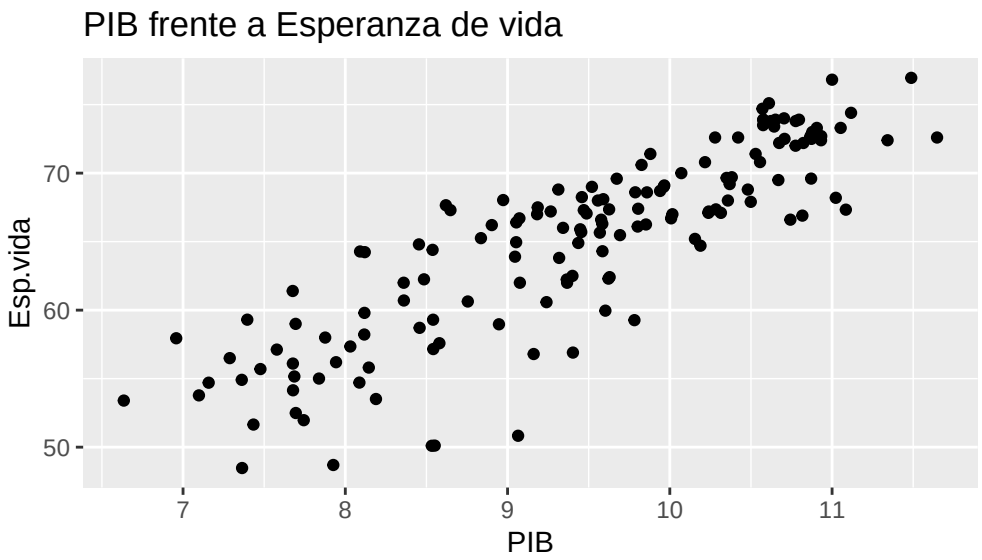

Figura 2: Representación del PIB frente a la esperanza de vida.

Por ejemplo en el PIB de un país y la esperanza de vida podemos suponer que tienen dependencia, más adelante veremos si es así, o por el contrario nos estamos equivocando.

3.3 Países más felices

Por lo general la calidad de vida será influyente en la felicidad, siendo factores como la pobreza y las guerras que afectan negativamente. En los países más felices 9 de los 10 del top se sitúan en Europa.

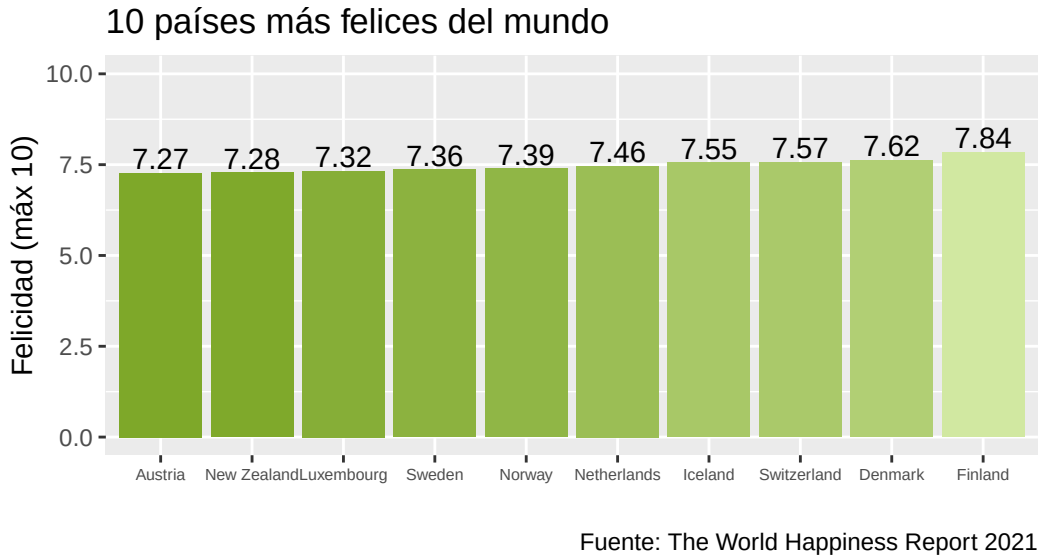

Figura 3: Top 10 países más felices

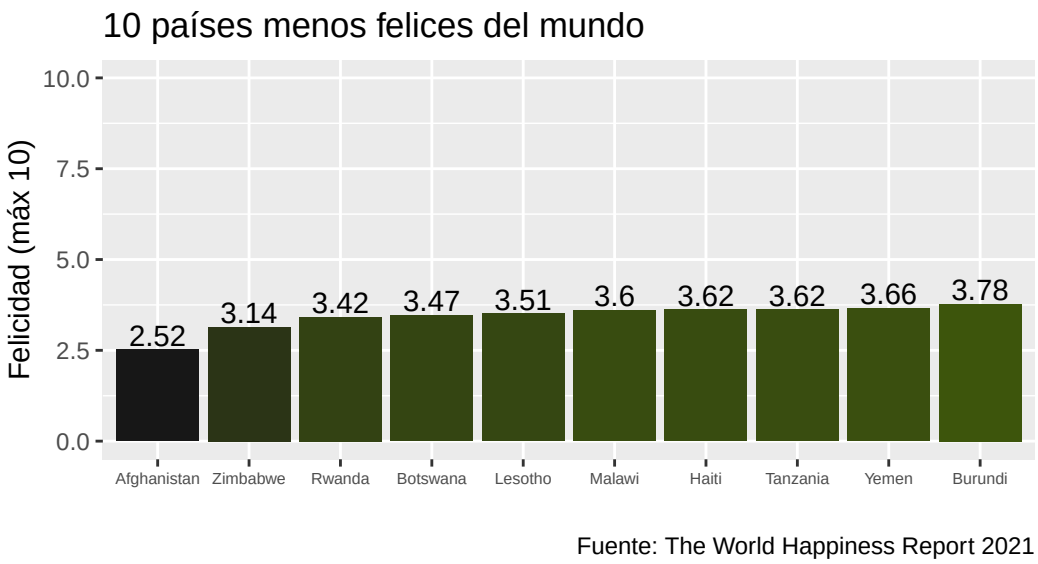

Figura 4: Top 10 países menos felices

3.4 Felicidad por region

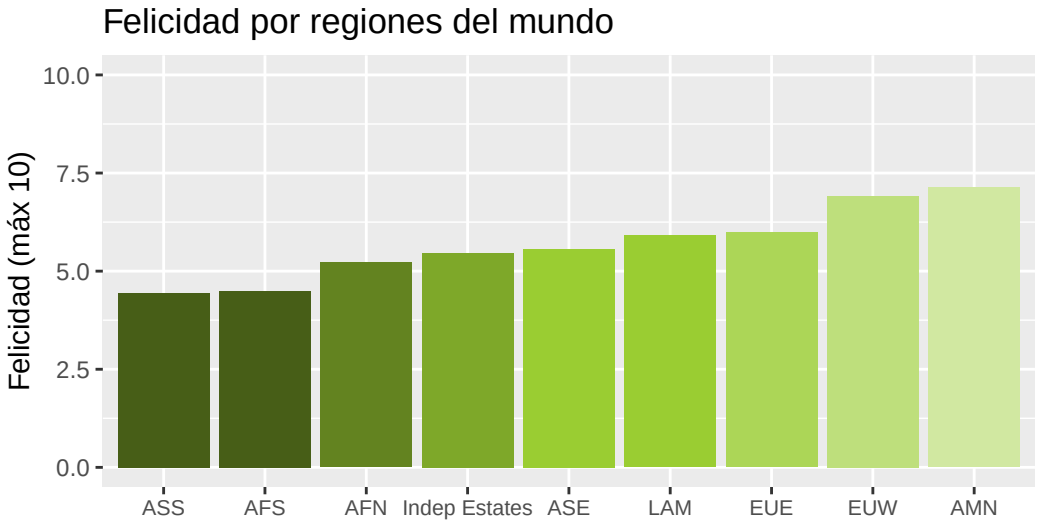

Figura 5: Gráfico de barras Felicidad por region.

4 PCA

4.1 Qué es un PCA

Como hemos visto en la figura 1 hay variables que dependen de otras en cierta medida, a esto se le llama **correlación** y adopta un valor de -1 a 1. Una alta correlación en nuestro datos significa que hay redundancia en nuestros datos, es decir tenemos información repetida, que con el cambio apropiado de variables nos permitirá guardar gran porcentage de la información con menos variables.

PCA (Análisis de componentes principales) es un método de análisis de datos multivariante que nos permite resumir y visualizar la información contenida en un gran conjunto de datos de variables cuantitativas.

4.2 Procedimiento

4.2.1 Correlación

El primer paso para realizar un PCA es ver si nuestras variables pueden estar relacionadas como hemos visto anteriormente, y a continuación crear una matriz de correlaciones como la siguiente:

Tabla 4: Correlación de las variables.

	Felicidad	PIB	Apoyo.social	Esp.vida	Libertad	Generosidad	Corrupcion
Felicidad	1.00	0.79	0.76	0.77	0.61	-0.02	-0.42
PIB	0.79	1.00	0.79	0.86	0.43	-0.20	-0.34
Apoyo.social	0.76	0.79	1.00	0.72	0.48	-0.11	-0.20
Esp.vida	0.77	0.86	0.72	1.00	0.46	-0.16	-0.36
Libertad	0.61	0.43	0.48	0.46	1.00	0.17	-0.40
Generosidad	-0.02	-0.20	-0.11	-0.16	0.17	1.00	-0.16
Corrupcion	-0.42	-0.34	-0.20	-0.36	-0.40	-0.16	1.00

Factores como la riqueza, la salud y el apoyo social son los más importantes para determinar la felicidad, mientras que la generosidad y la corrupción son las menos relacionadas.

Ahora calcularemos el índice de Kaiser-Meyer-Olkin (KMO), que compara los valores de correlaciones entre pares de variables. Si el índice adopta un valor proximo a 1 será viable realizar el PCA.

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = dfn)
## Overall MSA =  0.84
## MSA for each item =
##    Felicidad          PIB Apoyo.social     Esp.vida     Libertad  Generosidad
##         0.87         0.81         0.86         0.86         0.85         0.56
##   Corrupcion
##         0.78
```

El índice global es 0.84 por lo que vamos a realizar el análisis.

4.2.2 Cálculo PCA

Como ya lo hemos justificado procedemos a realizar los cálculos.

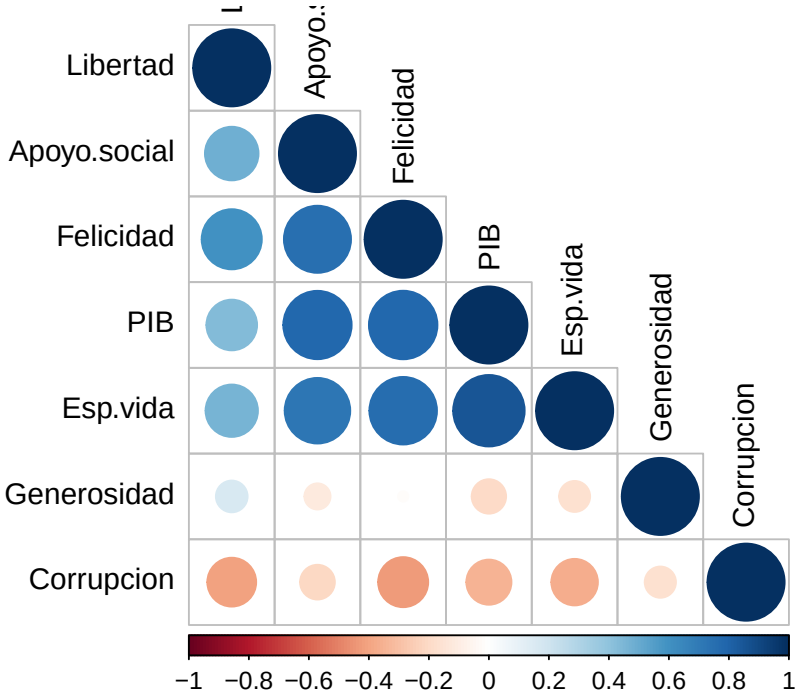

Figura 6: Correlación de las variables.

Tabla 5: Autovalores y varianza acumulada por cada componente.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.91	55.91	55.91
comp 2	1.29	18.42	74.33
comp 3	0.71	10.12	84.44
comp 4	0.52	7.41	91.85
comp 5	0.25	3.59	95.43
comp 6	0.19	2.76	98.20
comp 7	0.13	1.80	100.00

Lo que significa que con tan solo 3 componentes acumulamos un 84.44% de la varianza.

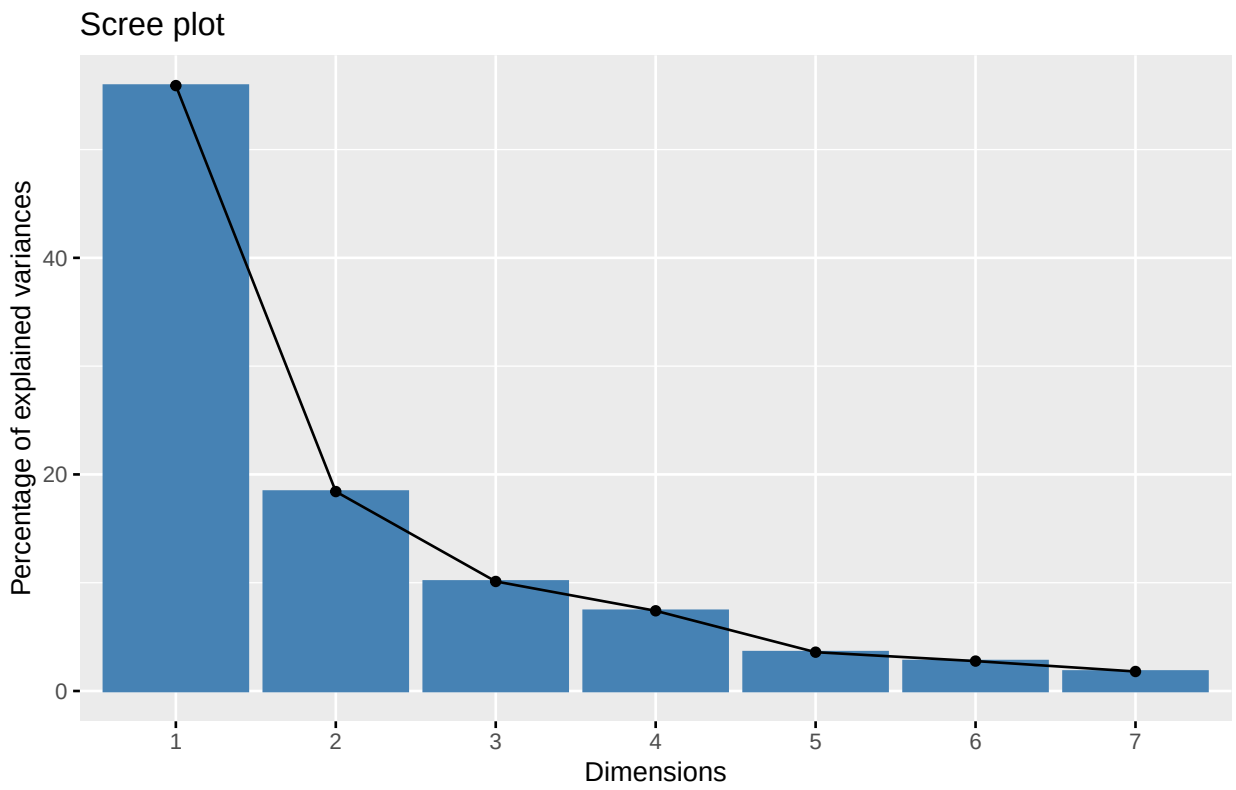

Figura 7: Gráfico de sedimentación de los autovalores

4.2.3 Representación gráfica. Interpretación

Procedemos a representar el peso que tiene cada variable de las originales en las nuevas componentes y los individuos en las dos primeras dimensiones principales.

La primera dimensión está relacionada con el bienestar del individuo, y la segunda diremos que tiene que ver con lo egoísta que es la sociedad.

Se sitúan a la izquierda países del Sur de África y Sur de Ásia, por media Norte de África, Latinoamérica y Este de Ásia, y finalmente a la derecha se sitúan los países de Norte de América y Europa.

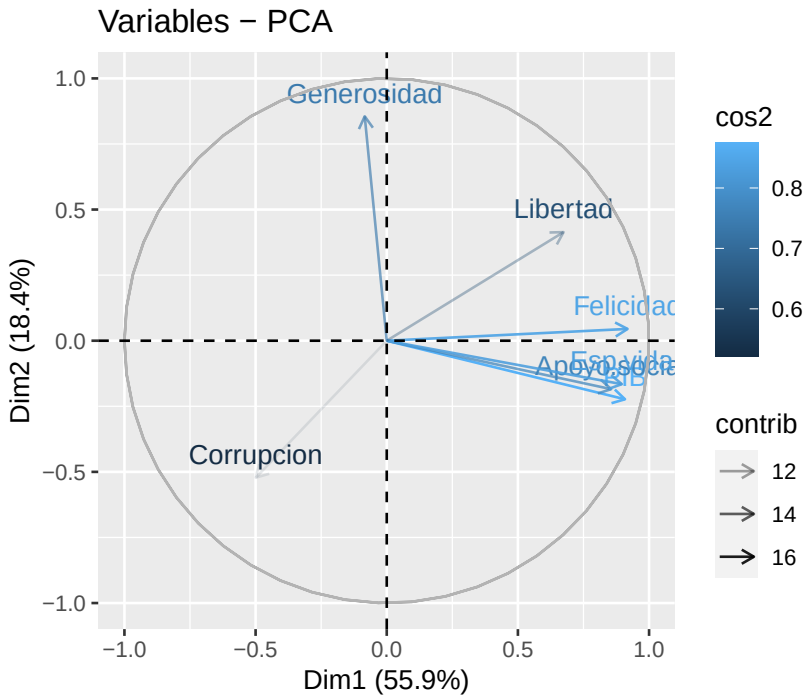

Figura 8: Representación variables PCA en dos dimensiones

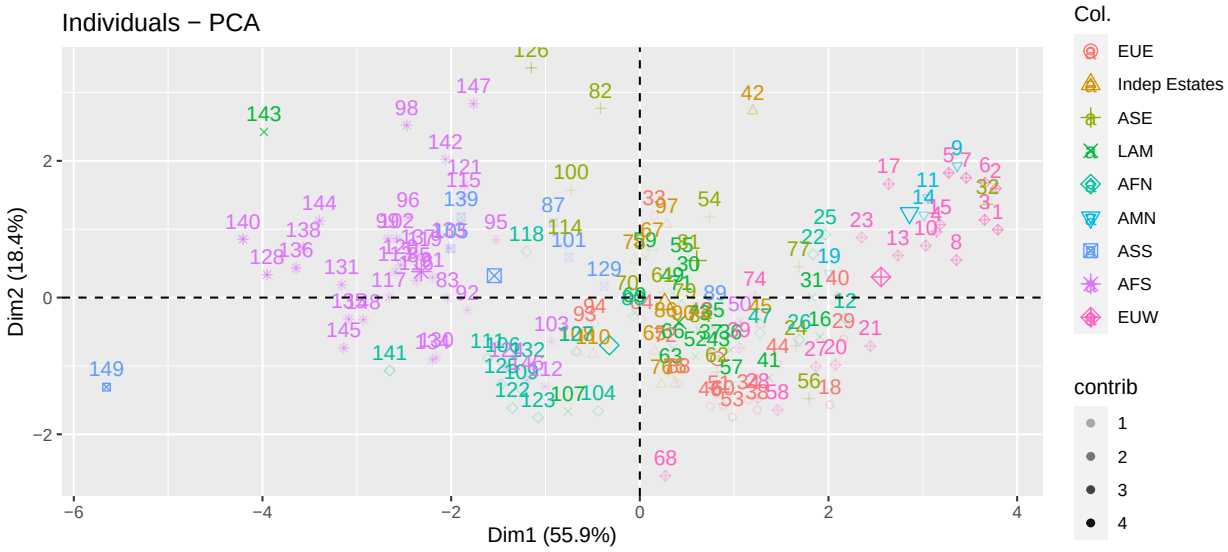

Figura 9: Representación de los individuos

5 Referencias y bibliografía

Adler, D., & Murdoch, D. (2021). *Rgl: 3D visualization using opengl.* https://CRAN.R-project.org/package=rgl

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for r.* https://shiny.rstudio.com/

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools.* https://CRAN.R-project.org/package=purrr

Husson, F., Josse, J., Le, S., & Mazet, J. (2020). *FactoMineR: Multivariate exploratory data analysis and data mining.* http://factominer.free.fr

Kassambara, A., & Mundt, F. (2020). *Factoextra: Extract and visualize the results of multivariate data analyses.* http://www.sthda.com/english/rpkgs/factoextra

Lê, S., Josse, J., & Husson, F. ois. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Luque-Calvo, P. L. (2017). *Escribir un trabajo fin de estudios con r markdown.* Disponible en http://destio.us.es/calvo.

M. Francisca Carreño Fructuoso, J. M. M. P., Fernando Pérez Sanz. (n.d.). *Curso online autónomo métodos de análisis de datos multivariantes.* 00R-team. https://argos.inf.um.es/mamutCola/index.html

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames.* https://CRAN.R-project.org/package=tibble

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research.* https://personality-project.org/r/psych/%20https://%0D%0A%20%20%20%20personality-project.org/r/psych-manual.pdf

team, O. (2021). *Latex documentation.* Overleaf-team. https://es.overleaf.com/learn

Thuleau, S., & Husson, F. (2020). *FactoInvestigate: Automatic description of factorial analysis.* http://factominer.free.fr/reporting/

Vaissie, P., Monge, A., & Husson, F. (2021). *Factoshiny: Perform factorial analysis from factominer with a shiny application.* http://factominer.free.fr/graphs/factoshiny.html

Wei, T., & Simko, V. (2021a). *Corrplot: Visualization of a correlation matrix.* https://github.com/taiyun/corrplot

Wei, T., & Simko, V. (2021b). *R package "corrplot": Visualization of a correlation matrix.* https://github.com/taiyun/corrplot

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2021). *Tidyr: Tidy messy data.* https://CRAN.R-project.org/package=tidyr

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2020). *Ggplot2: Create elegant data visualisations using the grammar of graphics.* https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation.* https://CRAN.R-project.org/package=dplyr

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research.* Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/%0D%0A%20%20%20%209781466561595

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.org/knitr/

Xie, Y. (2021). *Knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/

Yihui Xie, E. R., Christophe Dervieux. (2021). *R markdown cookbook*. https://bookdown.org/yihui/rmarkdown-cookbook/

Zhu, H. (2021). *KableExtra: Construct complex table with kable and pipe syntax*. https://CRAN.R-project.org/package=kableExtra

Índice de tablas

Índice de figuras