# EMBEBIMIENTOS MÉTRICOS, PARTE II: ANÁLISIS DE COMPONENTES PRINCIPALES, ANÁLISIS DE CONGLOMERADOS JERÁRQUICOS Y ANÁLISIS DE CONGLOMERADOS POR K-MEDIAS APLICADOS AL ANÁLISIS CANTONAL DE INDICADORES SOBRE LA NIÑEZ Y LA VEJEZ

**ISADORE NABI**

*I. INTRODUCCIÓN* 2
*II. MATERIALES Y MÉTODOS* 2
**II.I. Tipo de Investigación** 2
**II.II. Diseño de la Investigación** 2
**II.III. Alcance de la Investigación** 2
**II.IV. Muestra** 2
**II.I. Instrumentos** 3
*III. CONCEPTUALIZACIÓN* 3
**III.I. Librería 'factoextra'** 3
**III.II. Autovalores y Cargas Factoriales** 4
**III.III. Contribución** 5
**III.IV. Coseno Cuadrado (cos2)** 6
**III.V. Gráfica de Sedimentación** 7
**III.VI. Análisis de Componentes Principales** 7
**III.VII. Análisis de Conglomerados** 8
**III.VIII. Análisis de Conglomerados Jerárquicos** 8
**III.IX. Análisis de Conglomerados por K-Medias** 10
**III.X. Validación de Agrupamientos** 11
**III.XI. Establecer Semilla (set.seed)** 12
*IV. RESULTADOS EMPÍRICOS* 13
*V. CONCLUSIONES Y RECOMENDACIONES* 45
**III.I. Análisis de Componentes Principales** 45
**III.II. Análisis de Conglomerados Jerárquicos** 47
**III.III. Análisis de Conglomerados por K-Medias** 47

**III.IV. Elección del Mejor Método Descriptivo** **48**
***VI. REFERENCIAS*** ***49***

# I. INTRODUCCIÓN

El presente trabajo empírico estudia descriptiva y semidescriptivamente[1] diversos indicadores sociales de nutrición y cuidados médicos para 81 cantones de Costa Rica en el año 2019.

# II. MATERIALES Y MÉTODOS

## II.I. Tipo de Investigación

Investigación experimental.

## II.II. Diseño de la Investigación

Cuantitativa.

## II.III. Alcance de la Investigación

Descriptivo/exploratorio robustecido.

## II.IV. Muestra

Índices relativos al progreso social cantonal para los 81 cantones de Costa Rica de 2019 (INCAE Business School, 2022). Específicamente tales índices, relativos a nutrición y cuidados médicos, son:

[1] Por el uso de bootstrapping y valores p.

a) Prevalencia[2] de desnutrición infantil (1 mayor prevalencia; 5 menor prevalencia).

b) Prevalencia de sobrepeso y obesidad infantil (1 mayor prevalencia; 5 menor prevalencia).

c) Tasa de vacunación (cobertura bruta %).

d) Acceso a redes de cuido de personas de tercera edad (tasa de centros de cuido por 100,000 personas de la población objetivo). e) Mortalidad infantil (muertes por 1,000 nacimientos).

### II.I. Instrumentos

Análisis de Componentes Principales (PCA), Análisis de Conglomerados Jerárquicos y Análisis de Conglomerados por K-medias.

## III. CONCEPTUALIZACIÓN

### III.I. Librería 'factoextra'

Proporciona algunas funciones fáciles de usar para extraer y visualizar la salida de análisis de datos multivariados, incluidos 'PCA' (Análisis de componentes principales), 'CA' (Análisis de correspondencia), 'MCA' (Análisis de correspondencia múltiple), 'FAMD' (Análisis factorial de datos mixtos), funciones 'MFA' (Análisis factorial múltiple) y 'HMFA' (Análisis jerárquico factorial múltiple) de diferentes paquetes de R. También contiene funciones para simplificar algunos pasos del análisis de agrupamiento y proporciona una elegante visualización de datos basada en 'ggplot2'.

[2] En epidemiología, la prevalencia es la proporción de personas que sufren una enfermedad con respecto al total de la población en estudio.

**III.II. Autovalores y Cargas Factoriales**

Los elementos de los eigenvectores son las proporciones estimadas en que las variables explicativas (las columnas que contienen las ramas productivas del sistema económico) del comportamiento estadístico de los individuos participan en la conformación del nuevo espacio (conformado por las dimensiones, que en R aparecen como $Dim_i$) tomando en consideración las varianzas estimada sobre tales participaciones (que son los valores singulares correspondientes, expresados como la raíz cuadrada del eigenvalor correspondiente a cada eigenvector), concretamente realizando una estandarización de tales eigenvectores mediante su división entre su pertinente valor singular [véanse las discusiones (Cross Validated, 2015) y (Cross Validated, 2020)], el cual expresa la varianza de los elementos del eigenvector en cuestión.

No existe forma apriorística de determinar cuál variable explicativa es más importante en la conformación de cada una de las dimensiones; sin embargo, esta información puede conocerse de forma mediata y concreta a través de la revisión de los elementos de cada uno de los eigenvectores construidos para realizar el embebimiento métrico (existe un eigenvector por cada dimensión o componente principal generado) y determinando (mediante alguno de los múltiples criterios disponibles) cuál de los elementos del eigenvector (estos elementos son lo que en psicometría se conoce como *cargas factoriales*) tiene mayor participación dentro del mismo, puesto que cada carga factorial expresa la proporción en que cada una de las variables explicativas del comportamiento de los individuos participan en tal explicación; por supuesto, también se debe tomar en consideración su varianza (a través de los valores singulares) al momento de analizar las magnitudes de los componentes del eigenvector con el fin de determinar cuál (o cuáles) de las variables originales tiene (o tienen) mayor participación en las nuevas dimensiones conformadas.

Los eigenvalores obtenidos son el promedio de los valores contenidos dentro del eigenvector y, puesto que el valor esperado de una eigenfunción [o eigenvector, véase la discusión (Mathematics, 2012)] es igual a su promedio [véase (Ranganathan, 2022, pág. 2)], el eigenvalor puede verse como el valor esperado de las cargas factoriales expresadas en la forma antes señalada.

Los autovalores son magnitudes numéricas obtenidas como la derivada de la función evaluada en un punto[3] divididas entre la corrección de Bessel $(n-1)$, donde dicha corrección convierte a la varianza en una variación alrededor del origen). Los autovalores expresan la variación que el $PC_i$ al que correspondan explica del conjunto de datos estudiado.

**III.III. Contribución**

Como señalan (Abdi & Williams, 2010, pág. 5), el valor propio asociado a un componente es igual a la suma de las puntuaciones factoriales al cuadrado para este componente. Por lo tanto, la importancia de una observación para un componente puede obtenerse mediante la relación de la puntuación factorial al cuadrado de esta observación por el valor propio asociado con ese componente. Esta razón se llama la contribución de la observación al componente.

El valor de una contribución está entre 0 y 1 y, para un componente dado, la suma de las contribuciones de todas las observaciones es igual a 1. Cuanto mayor sea el valor de la contribución, más contribuye la observación al componente. Como señalan (Abdi & Williams, 2010, pág. 5), una heurística útil es basar la interpretación de un componente en las observaciones cuya contribución es mayor

[3] Es decir, son las pendientes de cada predictor/variable, donde dicha pendiente para cada predictor/variable está dividida entre la distancia máxima o hipotenusa, que en este contexto se expresa como la raíz cuadrada del eigenvalor correspondiente a cada eigenvector.

que la contribución promedio[4] (es decir, observaciones cuya contribución es mayor que $1/l$, donde $l$ es el $l - esimo$ componente principal); sin embargo, qué tan superior debe ser el valor de una contribución para considerarla como significativa dependerá, en última instancia, de las necesidades objetivas de la investigación. Las observaciones con contribuciones altas y signos diferentes se pueden oponer para ayudar a interpretar el componente porque estas observaciones representan los dos extremos de este componente. Las puntuaciones de los factores de las observaciones complementarias no se utilizan para calcular los valores propios y, por lo tanto, sus contribuciones generalmente no se calculan.

**III.IV. Coseno Cuadrado (cos2)**

Como señalan (Abdi & Williams, 2010, págs. 5-6), el coseno al cuadrado muestra la importancia de un componente para una observación dada. El coseno al cuadrado indica la contribución de un componente al cuadrado de la distancia de la observación al origen. Corresponde al cuadrado del coseno del ángulo del triángulo rectángulo formado con el origen, la observación y su proyección sobre la componente.

Se utiliza como medida de calidad en la representación de los individuos. Asumiendo que los individuos cuyo coseno cuadrado sea mayor a 0.50 son representativos, se concluiría que todos los individuos alrededor del origen estarían mal representados por el modelo PCA.

Sin embargo, como señala (Kassambara, PCA - Principal Component Analysis Essentials, 2017), al igual que en la valoración de la contribución, un coseno cuadrado superior a la media es una heurística útil para seleccionar las observaciones cuya calidad de representación es significativa. ¿Qué tan superior a

[4] Lo mismo se señala en (Kassambara, PCA - Principal Component Analysis Essentials, 2017), siendo esto el criterio de inclusión ("cutoff" en la fuente referida) de las observaciones de un componente.

la media debe considerarse como significativo?, esto dependerá de las necesidades objetivas de la investigación.

### III.V. Gráfica de Sedimentación

Gráfica de Sedimentación Señala (Minitab, 2019) que "(…) un gráfico de sedimentación muestra el número del componente principal versus su valor propio correspondiente. La gráfica de sedimentación ordena los valores propios desde el más grande hasta el más pequeño. Los valores propios de la matriz de correlación son iguales a las varianzas de los componentes principales." Esta gráfica sirve para seleccionar el número de componentes que se usarán con base en la magnitud (el "tamaño") de los valores característicos. El patrón ideal es una curva pronunciada, seguida de una inflexión y luego de una línea recta. Se deben utilizar los componentes en la curva pronunciada antes del primer punto que inicia la tendencia de línea.

### III.VI. Análisis de Componentes Principales

Como señala (Jollife, 2002), el PCA es una metodología estadística que busca reducir la dimensionalidad de un conjunto de datos que posee un gran número de variables interrelacionadas logrando retener tanto como sea posible la variabilidad actualmente presente en el conjunto de datos estudiado, puesto que esto implica que la distribución de tal conjunto de datos dentro del espacio muestral tras realizar la transformación ortogonal será isométricamente equivalente a la distribución del conjunto de datos original[5].

Algunos de los criterios más utilizados para seleccionar el número de nuevas dimensiones a utilizar son:

---

[5] Puede profundizarse sobre el fundamento formal en (Nabi, EMBEBIMIENTOS MÉTRICOS: ANÁLISIS DE COMPONENTES PRINCIPALES Y ANÁLISIS DE CONGLOMERADOS. SU INTERPRETACIÓN CONCEPTUAL, INTUICIÓN-LÓGICA GEOMÉTRICA, FORMALISMO MATEMÁTICO Y APLICACIONES EN RSTUDIO Y MINITAB, 2022, págs. 8-29).

1. Autovalor mayor a la unidad[6] (Universitat de Girona, 2009).
2. Varianza entre 70% y 90% (Everitt & Hothorn, 2011, pág. 71).
3. En la gráfica de sedimentación, el número de nuevas variables/dimensiones que están del punto donde se forma el codo hacia atrás (Minitab, 2019).

### III.VII. Análisis de Conglomerados

Como señalan (Hennig, Meila, Murtagh, & Rocci, 2016, pág. 2), formalmente, dado un conjunto de objetos, también llamado conjunto de datos, cuya estructura se define como $D = \{x_1, \dots, x_n\}$ y contiene n-ésimos puntos de datos u observaciones, la tarea de construir conglomerados con los elementos $x_i \in D$ consiste en agrupar tales elementos en K-ésimos conjuntos (y en función si tales conjuntos son disjuntos o no, es que se trata de clústeres particionales o clústeres jerárquicos o traslapados) pertenecientes a D, denotados estos como $C_1, \dots, C_k$. Un conglomerado es, como primera aproximación, cada una de las particiones obtenidas, esto es, $C = \{C_1, \dots, C_k\}$.

El número de clústeres es, en caso de utilizar en el análisis de conglomerados, dimensiones generadas mediante algún embebimiento métrico (por ejemplo, el realizado vía PCA en esta investigación), la cantidad de conglomerados bajo los cuales se agruparán las observaciones contenidas en las dos dimensiones generadas por el embebimiento métrico de tipo PCA.

### III.VIII. Análisis de Conglomerados Jerárquicos

Como señala (Bock, 2021), el agrupamiento jerárquico comienza tratando cada observación como un grupo separado. Luego, ejecuta repetidamente los siguientes dos pasos:

1. Identifica los dos clústeres que están más cerca entre sí.

---

[6] Conocido como criterio de Kaiser, elaborado en 1960.

2. Fusiona los dos clústeres más similares. Este proceso iterativo continúa hasta que todos los clústeres se fusionan.

El método de validación de conglomerados jerárquicos utilizado aquí es el proporcionado por la función 'pvclust', que permite robustecer los resultados con bootstrapping y la generación de valores p. Con la sintaxis 'pvpick', de la misma librería que 'pvclust', se encuentran los clústeres con valores p relativamente altos/bajos (según los criterios preestablecidos por el investigador, tal como se verá en la sección relativa a la estimación de los conglomerados jerárquicos significativos). Los componentes de la sintaxis 'pvpick' son:

1. "alpha" es el valor mínimo de una magnitud a partir de cual se produce un efecto determinado. "pv" es una cadena de caracteres que especifica el valor p que se utilizará. Debe ser uno de "si", "au" y "bp", correspondiente a SI, valor p de AU y valor de BP, respectivamente.

2. "tipo" es uno de "geq", "leq", "gt" o "lt". Si se especifica "geq", se devuelven o muestran los clústeres con un valor p mayor o igual al umbral dado por "alfa". Del mismo modo, "leq" significa menor que o igual, "gt" mayor que y "lt" menor que el valor umbral. El valor predeterminado es "geq". c) "max.only" es un valor lógico. Si algunos de los clústeres con valores p altos/bajos tienen una relación de inclusión, solo se devuelve (o muestra) el clúster más grande cuando max.only=TRUE.

Como se señala en (stack overflow, 2017), los métodos utilizados por 'pvclust' tienen su origen en el análisis filogenético en el que intenta agrupar individuos de diferentes especies en función de su ADN u otras características. Sin embargo, el caso filogenético o genómico es un poco peculiar. En esta situación, las observaciones/muestras son los individuos (fuente del ADN) y las variables/características/descriptores son, por ejemplo, los genes o su nivel de expresión.

En tal caso, tiene sentido agrupar a los individuos en función de sus genes y probar la estabilidad de los grupos arrancando los genes porque los genes son múltiples y es razonable volver a muestrear al azar dentro del conjunto de genes.

La interpretación de BP (probabilidad bootstrapping) es bastante sencilla. 'pvclust' remuestreará las columnas del conjunto de datos, que son aquí los indicadores analizados[7], es decir, que algunas columnas se eliminarán y otras aparecerán más de una vez (por la naturaleza de la técnica boostrapping, que genera isometrías al conjunto de datos original, de ahí que se genere un tensor). Luego, volverá a calcular la distancia euclidiana entre las filas (cantones) y construirá el dendrograma. Repetirá esto nboot*length(r) veces [length(r) = 10 por defecto] y para cada ejecución verificará para cada grupo si el mismo conjunto de cantones está presente en un grupo del conjunto de datos de arranque.

**III.IX. Análisis de Conglomerados por K-Medias**

Como señala (Nabi, Sobre los Estimadores de Bayes, el Análisis de Grupos y las Mixturas Gaussianas, 2020, págs. 67-68), un análisis de conglomerados por K-medias es un método de cuantización vectorial[8] que busca particionar $n$ observaciones en $k$ grupos en los que cada observación pertenece al grupo con la media más cercana (que representa el centro o centroide del grupo desde la perspectiva geométrica), siendo precisamente tales medias el vector prototipo de cada k-ésimo grupo. Esto pertenece a las técnicas de clasificación del aprendizaje automático.

---

[7] Para este caso concreto indicadores compuestos, puesto que las nuevas dimensiones/variables generadas vía el embebimiento métrico PCA (que son las variables que se utilizan para construir los clústeres) son conceptualmente hablando indicadores compuestos.

[8] Técnica que permite modelar las funciones de densidad de probabilidad de un conjunto de datos usando la función de distribución acumulada de los vectores prototipo (que es el elemento del espacio muestral que representa a un grupo de elementos dentro de dicho espacio).

**III.X. Validación de Agrupamientos**

Como se señala en (Kassambara, Determining The Optimal Number Of Clusters: 3 Must Know Methods , 2022), determinar el número óptimo de conglomerados en un conjunto de datos es un tema fundamental en el agrupamiento de particiones, como el agrupamiento de k-medias, que requiere que el usuario especifique el número de conglomerados k que se generarán.

Estos métodos incluyen métodos directos y métodos de prueba estadísticos.

Los métodos directos consisten en optimizar un criterio, como las sumas de cuadrados dentro de un grupo o la silueta promedio; entre estos dos métodos, como indica (Kumar, 2020), es más robusto el método de la silueta. Los métodos correspondientes se denominan métodos de codo y silueta, respectivamente. Por su parte, los métodos estadísticos de prueba consisten en comparar la evidencia contra la hipótesis nula. Un ejemplo es el método del estadístico de la brecha. La desventaja de los métodos de codo y silueta promedio es que solo miden una característica de agrupación global. Un método más sofisticado es usar la estadística de brecha que proporciona un procedimiento estadístico para formalizar la heurística (la técnica de indagación y de descubrimiento) de los métodos codo y silueta para estimar el número óptimo de grupos.

Además de los métodos estadísticos de codo (Elbow), silueta y brecha, se han publicado más de treinta índices y métodos para identificar el número óptimo de grupos. En esta investigación se utilizará una función que permite calcular todos estos 30 índices y decidir el número óptimo de conglomerados utilizando la "regla de la mayoría". La regla de la mayoría es una regla de decisión que selecciona el número óptimo de clústeres escogido por la mayoría de los 30 criterios utilizados, concluyendo que el mejor número de clústeres es aquel por el que más criterios de entre los 30 utilizados se decantan.

Para el caso del método de la silueta, la puntuación de la silueta se encuentra dentro del rango [-1, 1]. La puntuación de silueta de 1 significa que los conglomerados son muy densos y bien separados. La puntuación de 0 significa que los clústeres se superponen (no están bien separados). La puntuación de menos de 0 significa que los datos pertenecientes a los clústeres pueden ser erróneos/incorrectos (mal separados).

**III.XI. Establecer Semilla (set.seed)**

Como se señala en (R CODER, 2022), cuando se generan "números aleatorios" en R, en realidad se están generando números pseudoaleatorios. Estos números se generan con un algoritmo que requiere una semilla o valor inicial para inicializar. Que los números sean pseudo-aleatorio en lugar de "aleatorios puros" (la aleatoriedad objetivamente no existe, es decir, no es una característica intrínseca de la Naturaleza, sino únicamente un recurso cognitivo que el intelecto humano utiliza para aproximarse a la verdad) significa que, si se conoce la semilla y el generador de tales números, se puede predecir (y reproducir) la salida. La relevancia de configurar una semilla (por ejemplo, con la sintaxis set.seed que aquí se utilizará), cómo funciona set.seed y cómo configurar o desactivar la semilla, reside en la reproductibilidad de los resultados de la investigación científica. Establecer una semilla en R u otro programa computacional significa inicializar un generador de números pseudoaleatorios. La mayoría de los métodos de simulación en Estadística requieren la posibilidad de generar números pseudoaleatorios que imiten las propiedades de generaciones independientes de una distribución uniforme en el intervalo (0,1). Para obtener estas secuencias de números pseudoaleatorios, se necesita un algoritmo recursivo llamado Generador de Números Aleatorios (RNG), por ejemplo, la sintaxis "rnorm(X)". Sin embargo, si se ejecuta esta sintaxis u otra, sin instrucciones adicionales, se obtendrá una salida diferente. Esto implica que el código no es reproducible, porque no conoce la semilla que R usó para generar esa secuencia. Así, establecer una semilla en R se

usa para: 1) reproducir el mismo resultado de los estudios de simulación, 2) ayudar a depurar el código cuando se trata de números pseudoaleatorios. Al llamar a una función de generación de números aleatorios, la salida depende de los valores de ".Random.seed", que cambia después de ejecutar estas funciones. Si almacena el valor de ".Random.seed", se puede obtener el estado inicial actual. En consecuencia, en caso de que se desee generar los mismos números una K cantidad de veces, se debe configurar la misma semilla K veces, por ejemplo, set.seed(123).

## IV. RESULTADOS EMPÍRICOS

### *4.1. ANÁLISIS DESCRIPTIVO GENERAL*

```
##Cargando el conjunto de datos tomado del INCAE, quienes a su vez señalan en s
u página web (https://www.incae.edu/es/clacds/proyectos/indice-de-progreso-soci
al-cantonal-2019.html) que los construyen Utilizando 53 indicadores sociales y a
mbientales de fuentes públicas; sin embargo, no señalan las fuentes específicas de c
ada uno de los indicadores.
library(readxl)
DATOSCLUSTER <- read_excel("D:/Carpeta de Estudio/Mis Códigos en R/DAT
OSCLUSTER.xlsx",
   col_types = c("text", "numeric", "numeric",
      "numeric", "numeric", "numeric"))
View(DATOSCLUSTER)
DATA <- DATOSCLUSTER[,2:6]
summary(DATA)
```

```
##  Desnutrición infantil Tasa de vacunación Sobrepeso y obesidad infantil
##  Min.   :1.000        Min.   : 70.66     Min.   :1.000
##  1st Qu.:3.000        1st Qu.: 90.49     1st Qu.:2.000
##  Median :4.000        Median : 97.22     Median :3.000
##  Mean   :3.444        Mean   : 97.02     Mean   :2.679
##  3rd Qu.:4.000        3rd Qu.:102.96     3rd Qu.:3.000
##  Max.   :4.000        Max.   :135.71     Max.   :4.000
##  Acceso a redes de cuido personas de tercera edad Mortalidad Infantil
##  Min.   : 0.000                                   Min.   : 0.000
##  1st Qu.: 0.000                                   1st Qu.: 6.219
##  Median : 8.664                                   Median : 7.678
##  Mean   : 6.437                                   Mean   : 8.251
##  3rd Qu.:11.396                                   3rd Qu.: 9.950
##  Max.   :16.162                                   Max.   :22.642
```

***##En promedio, un 80% de los infantes muestreados no sufre de desnutrición, la tasa de vacunación es en promedio elevada (97.02%), es preocupante el valor promedio del indicador de sobrepeso y obesidad infantil (2.679, que indica que hay que mejorar los hábitos alimenticios, la calidad de alimentación y la cultura hacia el deporte u otros tipos de actividad física para la población infantil), nuestros ancianos y ancianas no reciben el cuido adecuado (en promedio apenas un 6.437% de cada 100,000 ancianos) y la tasa de mortalidad infantil es en promedio baja (de 8.251%, sin embargo, puede y debe ser menor, lo que requiere del acompañamiento de políticas públicas en materia de acceso a condiciones materiales de existencia óptimas y mejoramiento de las instituciones de salud). Salvo la tasa de vacunación, que parecería distribuirse normalmente en una primera mirada (porque su media y su mediana son muy similares), no parece que ninguna otra variable siga esa misma distribución.***

***##Es recomendable escalar (no debe confundirse con estandarizar) la base de datos porque la técnica de PCA combina las variables para generar las nuevas dimensiones en las cuales se realiza el embebimiento métrico del espacio muestral original y para que esta mezcla de variables se realice de la forma más adecuada, dado que en tal combinación se mezclarán sus escalas de medida originales, es necesaria la escalación de las mismas bajo una escala común, que para el caso de la sintaxis scale(X) es el número de columnas de X (véase la documentación de R relativa a dicha sintaxis).***

```
sDATA <- scale(DATA)
View(sDATA)
summary(sDATA)
```

```
##  Desnutrición infantil Tasa de vacunación Sobrepeso y obesidad infantil
##  Min.   :-2.8709       Min.   :-2.6125    Min.   :-2.0130
##  1st Qu.:-0.5220       1st Qu.:-0.6473    1st Qu.:-0.8141
##  Median : 0.6525       Median : 0.0199    Median : 0.3848
##  Mean   : 0.0000       Mean   : 0.0000    Mean   : 0.0000
##  3rd Qu.: 0.6525       3rd Qu.: 0.5889    3rd Qu.: 0.3848
##  Max.   : 0.6525       Max.   : 3.8361    Max.   : 1.5838
##  Acceso a redes de cuido personas de tercera edad Mortalidad Infantil
##  Min.   :-1.0887                                  Min.   :-1.8518
##  1st Qu.:-1.0887                                  1st Qu.:-0.4561
##  Median : 0.3768                                  Median :-0.1288
##  Mean   : 0.0000                                  Mean   : 0.0000
##  3rd Qu.: 0.8389                                  3rd Qu.: 0.3813
##  Max.   : 1.6450                                  Max.   : 3.2296
```

*#4.2. GENERACIÓN DE LOS COMPONENTES PRINCIPALES Y GRÁFICA DE SEDIMENTACIÓN*

```
library("FactoMineR")
res.pca <- PCA(X = sDATA, scale.unit = TRUE,graph = FALSE) #Aquí sí se estandari
za el conjunto de datos (media nula y varianza unitaria)
summary(res.pca)

##
## Call:
## PCA(X = sDATA, scale.unit = TRUE, graph = FALSE)
##
##
## Eigenvalues
##                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
## Variance              1.552   1.188   0.901   0.731   0.628
## % of var.            31.044  23.759  18.028  14.616  12.554
## Cumulative % of var. 31.044  54.803  72.830  87.446 100.000
##
## Individuals (the 10 first)
##                                      Dist    Dim.1    ctr
## 1                                  |  1.936 | -0.539  0.231
## 2                                  |  1.692 |  0.340  0.092
## 3                                  |  1.554 | -0.373  0.110
## 4                                  |  2.459 |  0.352  0.099
## 5                                  |  3.248 |  3.030  7.300
## 6                                  |  1.586 |  0.816  0.530
## 7                                  |  2.213 | -0.077  0.005
## 8                                  |  1.761 | -0.442  0.155
## 9                                  |  1.757 | -0.296  0.070
## 10                                 |  1.470 |  1.388  1.533
##                                     cos2    Dim.2    ctr   cos2
## 1                                  0.077 | -1.138  1.346  0.346
## 2                                  0.040 | -1.358  1.917  0.644
## 3                                  0.057 | -1.256  1.639  0.653
## 4                                  0.021 | -1.494  2.321  0.369
## 5                                  0.870 |  0.923  0.885  0.081
## 6                                  0.265 | -1.052  1.150  0.440
## 7                                  0.001 |  0.557  0.322  0.063
## 8                                  0.063 | -1.168  1.417  0.439
## 9                                  0.028 | -1.450  2.184  0.680
## 10                                 0.892 | -0.260  0.070  0.031
##                                     Dim.3    ctr   cos2
## 1                                  | -0.723  0.715  0.139 |
## 2                                  |  0.420  0.242  0.062 |
## 3                                  | -0.490  0.329  0.100 |
```

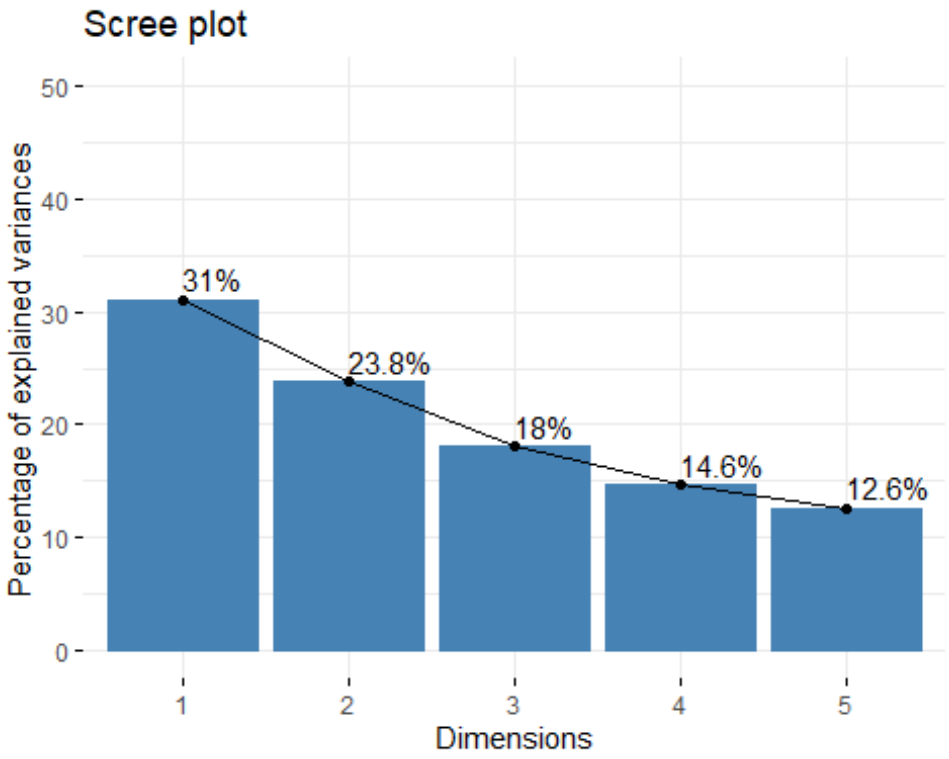
```
## 4                                    |  1.329  2.420  0.292 |
## 5                                    | -0.334  0.153  0.011 |
## 6                                    |  0.542  0.402  0.117 |
## 7                                    |  1.764  4.260  0.635 |
## 8                                    | -0.627  0.538  0.127 |
## 9                                    | -0.767  0.805  0.190 |
## 10                                   |  0.177  0.043  0.014 |
##
## Variables
##                                       Dim.1    ctr   cos2
## Desnutrición infantil                          |  0.498 15.961  0.248 |
## Tasa de vacunación                             | -0.289  5.389  0.084 |
## Sobrepeso y obesidad infantil                  |  0.472 14.377  0.223 |
## Acceso a redes de cuido personas de tercera edad | -0.683 30.032  0.466 |
## Mortalidad Infantil                            |  0.729 34.242  0.531 |
##                                     Dim.2    ctr   cos2    Dim.3
## Desnutrición infantil                         -0.603 30.625  0.364 |  0.318
## Tasa de vacunación                             0.527 23.423  0.278 |  0.790
## Sobrepeso y obesidad infantil                  0.677 38.586  0.458 | -0.276
## Acceso a redes de cuido personas de tercera edad  0.102  0.873  0.010 | -0.316
## Mortalidad Infantil                            0.278  6.493  0.077 | -0.021
##                                      ctr   cos2
## Desnutrición infantil                       11.201  0.101 |
## Tasa de vacunación                          69.201  0.624 |
## Sobrepeso y obesidad infantil                8.470  0.076 |
## Acceso a redes de cuido personas de tercera edad 11.082  0.100 |
## Mortalidad Infantil                         0.047  0.000 |
```

```
library(factoextra)
```

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

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

```
autovalores = get_eigenvalue(res.pca)
fviz_screeplot(res.pca,addlabels = TRUE,ylim = c(0,50))
```




*#4.3. SELECCIÓN DE COMPONENTES PRINCIPALES*

res.pca$var$contrib *#Contribución de variables a cada nueva dimensión conformada (en este caso, la sintaxis conformó automáticamente 5 dimensiones).*

```
##                                        Dim.1      Dim.2
## Desnutrición infantil                 15.960901 30.6247998
## Tasa de vacunación                     5.389067 23.4233013
## Sobrepeso y obesidad infantil         14.376567 38.5864091
## Acceso a redes de cuido personas de tercera edad 30.031653  0.8727096
## Mortalidad Infantil                   34.241812  6.4927802
##                                        Dim.3      Dim.4
## Desnutrición infantil                 11.20081182 21.8420137
## Tasa de vacunación                    69.20094430  1.8010394
## Sobrepeso y obesidad infantil          8.46959122  0.0415798
## Acceso a redes de cuido personas de tercera edad 11.08173133 57.6509581
## Mortalidad Infantil                    0.04692134 18.6644090
##                                        Dim.5
## Desnutrición infantil                 20.371474
## Tasa de vacunación                     0.185648
## Sobrepeso y obesidad infantil         38.525853
## Acceso a redes de cuido personas de tercera edad  0.362948
## Mortalidad Infantil                   40.554077
```

res.pca$var$contrib

```
##                                                 Dim.1      Dim.2
## Desnutrición infantil                       15.960901 30.6247998
## Tasa de vacunación                           5.389067 23.4233013
## Sobrepeso y obesidad infantil               14.376567 38.5864091
## Acceso a redes de cuido personas de tercera edad 30.031653  0.8727096
## Mortalidad Infantil                         34.241812  6.4927802
##                                                 Dim.3      Dim.4
## Desnutrición infantil                       11.20081182 21.8420137
## Tasa de vacunación                          69.20094430  1.8010394
## Sobrepeso y obesidad infantil                8.46959122  0.0415798
## Acceso a redes de cuido personas de tercera edad 11.08173133 57.6509581
## Mortalidad Infantil                          0.04692134 18.6644090
##                                                 Dim.5
## Desnutrición infantil                       20.371474
## Tasa de vacunación                           0.185648
## Sobrepeso y obesidad infantil               38.525853
## Acceso a redes de cuido personas de tercera edad  0.362948
## Mortalidad Infantil                         40.554077
```

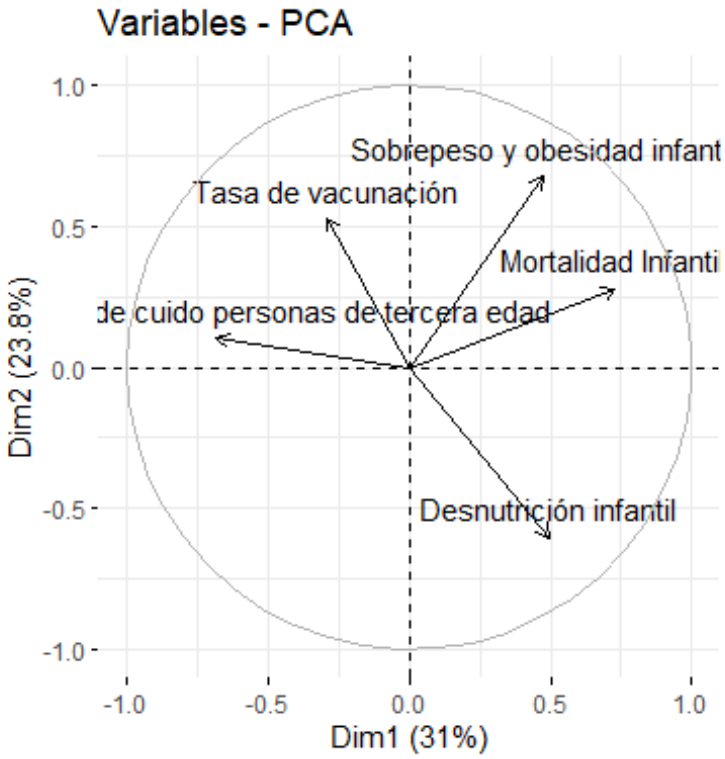
***##El criterio 1 sólo lo cumplen las primeras dos dimensiones, mientras que en el caso del tercer criterio, no existe un patrón claro de en qué número de dimensión se forma el codo, pero parecería que las candidatas estarían entre la dimensión 2 y 3. Finalmente, con relación al segundo criterio, para que la varianza explicada por el conjunto de dimensiones sea entre el 70% y el 90% se deberían tomar al menos 3 dimensiones, sin embargo, la tercera dimensión está significativamente por debajo de la magnitud deseada para su autovalor (más de un 10%, ya que el autorvalor debe ser mayor a la unidad y el autovalor de la dimensión 3 es de 0.90), además del hecho que esta dimensión 3 está conformada fundamentalmente por "tasa de vacunación" (aporta un aproximadamente el 70% a la dimensión) y tasa de vacunación ya tiene una participación importante en la dimensión 2, como se mostró anteriormente y se especificará a continuación. Por tanto, se trabajará únicamente con dos dimensiones.***

***##Así, las nuevas variables pueden representarse como se muestea a continuación. Puesto que la DIM1 está principalmente conformada por "Mortalidad infantil" (34.24%) y "Acceso a redes de cuido de personas de tercera edad" (30.03%), la nueva variable puede llamarse "Atención médica para preservación de la vida". Por otro lado, la DIM2 es está conformada principalmente por "Desnutrición infantil" (30.62%), "Tasa de vacunación" (23.48%) y "Sobrepeso y obesidad infantil" (38.57%), la nueva variable puede llamarse "Calidad de alimentación e inoculación de niños". En la elaboración anterior, se ha procurado que en las dimensiones conformadas no se repitan las variables.***

*#4.4. VARIABLES*

*#4.4.1. RELACIÓN ENTRE VARIABLES*

```
fviz_pca_var(res.pca,col.var = "black")
```




***##Las variables mejor representadas (cerca de la circunferencia) son "sobrepeso y obesidad infantil", "mortalidad infantil" y "desnutrición infantil", las más alejadas de la circunferencia (y, por tanto, más cercanas al centro) no están tan bien representadas como las variables antes mencionadas. Adicionalmente, con dos dimensiones se representa el 54.8% de variabilidad. Con relación a las correlaciones, parece existir una correlación inversa entre "tasa de vacunación" y "mortalidad infantil" (se forma un ángulo mayor a 90 grados entre ellas); lo mismo ocurre, aunque con correlación directa, entre "tasa de vacunación" y "acceso a redes de cuido de personas de la tercera edad"; de igual forma con "sobrepeso y obesidad infantil" con "mortalidad infantil" y entre "mortalidad infantil" y "desnutrición infantil". Finalmente, existe una correlación inversa entre "sobrepeso y obesidad infantil" y "desnutrición infantil" (su ángulo es mayor a 90 grados). Las demás correlaciones (como la correlación inversa que pareciera existir entre "tasa de vacunación" y "desnutrición infantil", salvo que se añadiera una variable latente que permita vincularlas orgánicamente desde algún marco teórico) no tienen sentido desde la lógica ni desde la medicina, por eso no se especifican.***

***##La interpretación anterior tiene como fundamento una lógica geométrica, ya que dos vectores son linealmente independientes si su producto escalar es nulo y esto, a nivel de dos dimensiones se expresa como una relación de perpendicularidad (que entre sus puntos de tangencia se forma un ángulo de 90 grados) o, de forma general***

***(n-dimensional), una relación de ortogonalidad.***

*#4.4.2. CONTRIBUCIÓN DE VARIABLES*

```
res.pca$var$contrib
```

```
##                                                 Dim.1      Dim.2
## Desnutrición infantil                       15.960901 30.6247998
## Tasa de vacunación                           5.389067 23.4233013
## Sobrepeso y obesidad infantil               14.376567 38.5864091
## Acceso a redes de cuido personas de tercera edad 30.031653  0.8727096
## Mortalidad Infantil                         34.241812  6.4927802
##                                                 Dim.3      Dim.4
## Desnutrición infantil                       11.20081182 21.8420137
## Tasa de vacunación                          69.20094430  1.8010394
## Sobrepeso y obesidad infantil                8.46959122  0.0415798
## Acceso a redes de cuido personas de tercera edad 11.08173133 57.6509581
## Mortalidad Infantil                          0.04692134 18.6644090
##                                                 Dim.5
## Desnutrición infantil                       20.371474
## Tasa de vacunación                           0.185648
## Sobrepeso y obesidad infantil               38.525853
## Acceso a redes de cuido personas de tercera edad  0.362948
## Mortalidad Infantil                         40.554077
```

```
mean(res.pca$var$contrib)
```

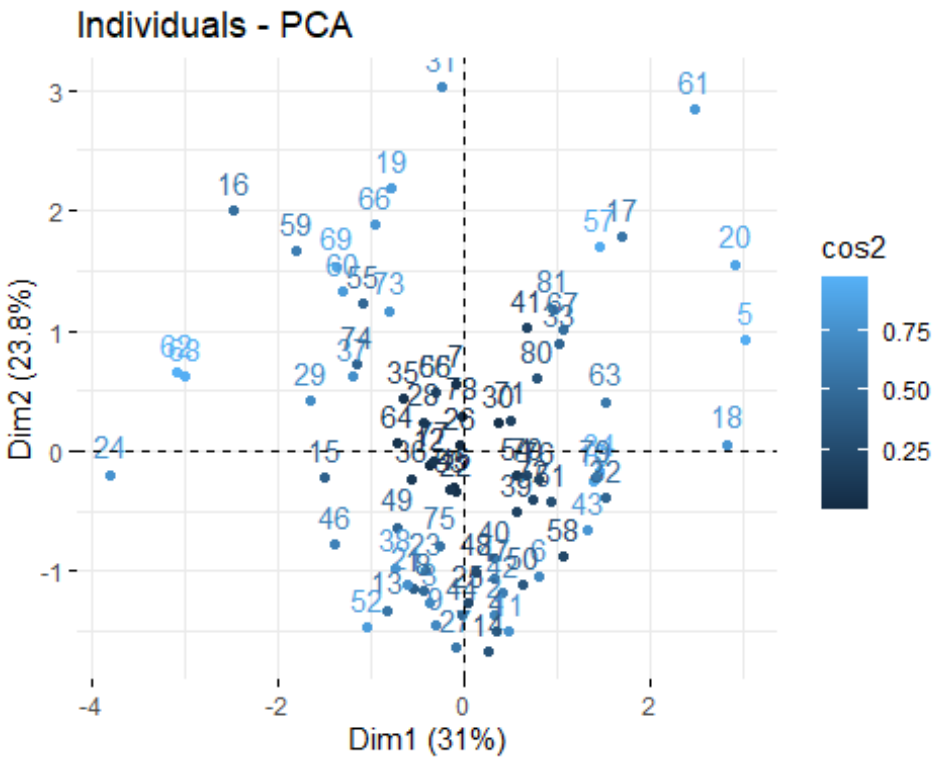
```
## [1] 20
```

*#4.4.2. CALIDAD DE REPRESENTACIÓN DE VARIABLES*

```
res.pca$var$cos2
```

```
##                                                 Dim.1      Dim.2
## Desnutrición infantil                       0.2477436 0.36380412
## Tasa de vacunación                          0.0836486 0.27825467
## Sobrepeso y obesidad infantil               0.2231517 0.45838323
## Acceso a redes de cuido personas de tercera edad 0.4661485 0.01036726
## Mortalidad Infantil                         0.5314982 0.07713031
##                                                 Dim.3       Dim.4
## Desnutrición infantil                       0.1009628310 0.1596227246
## Tasa de vacunación                          0.6237693622 0.0131621021
## Sobrepeso y obesidad infantil               0.0763439223 0.0003038676
## Acceso a redes de cuido personas de tercera edad 0.0998894532 0.4213166016
## Mortalidad Infantil                         0.0004229436 0.1364006019
##                                                 Dim.5
## Desnutrición infantil                       0.127866696
```

```
## Tasa de vacunación                                  0.001165266
## Sobrepeso y obesidad infantil                       0.241817236
## Acceso a redes de cuido personas de tercera edad 0.002278135
## Mortalidad Infantil                                 0.254547899

mean(res.pca$var$cos2)

## [1] 0.2

#4.5. INDIVIDUOS
#4.5.1. LOCALIZACIÓN ESPACIAL BIDIMENSIONAL DE INDIVIDUOS
fviz_pca_ind(res.pca, geom.ind =c("point", "text"), col.ind = "cos2", axes = c(1, 2))
```




```
#4.5.2. CONTRIBUCIÓN DE INDIVIDUOS
res.pca$ind$contrib #Contribución de individuos a cada nueva dimensión conformada

##         Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
## 1  2.306629e-01 1.346372449 7.153134e-01 2.677012109 1.057626e-01
## 2  9.210676e-02 1.917098652 2.421730e-01 1.221260828 6.326514e-03
## 3  1.104520e-01 1.638982442 3.292554e-01 0.762208795 1.497945e-02
## 4  9.878584e-02 2.321016554 2.420157e+00 0.167328104 3.581819e+00
## 5  7.300232e+00 0.885310796 1.525426e-01 0.446431010 2.843537e-01
## 6  5.301758e-01 1.150273381 4.021944e-01 0.337714673 4.855390e-01
## 7  4.746120e-03 0.321917828 4.260118e+00 0.866374238 1.885608e+00
## 8  1.554571e-01 1.416570790 5.384673e-01 1.869223936 8.658882e-02
```

## 9  6.972363e-02 2.183616873 8.053066e-01 0.519028092 8.557650e-03
## 10 1.532695e+00 0.070266749 4.268330e-02 0.197417134 3.330194e-02
## 11 1.827170e-01 2.337375631 2.352600e-02 1.229618738 9.687714e-04
## 12 1.014893e-01 0.015936860 3.090966e+00 2.282328989 5.741859e+00
## 13 5.380693e-01 1.821284399 9.821033e-01 1.948945321 1.112787e-02
## 14 5.121320e-02 2.889378745 4.980007e+00 2.838489706 1.084939e+00
## 15 1.774751e+00 0.051849586 1.502402e+00 3.383751156 1.338464e-01
## 16 4.889481e+00 4.148039239 9.997243e+00 0.008555264 3.874374e+00
## 17 2.298290e+00 3.285487526 7.512371e-02 7.388838897 8.608389e-02
## 18 6.371242e+00 0.002847846 2.798102e+00 0.003051676 2.712448e-02
## 19 4.952925e-01 4.971126316 1.028973e+00 0.076639024 4.186804e-01
## 20 6.839104e+00 2.505443545 5.189724e-01 1.063145205 5.218753e-01
## 21 2.984205e-01 1.292731120 5.107309e-02 0.805773863 6.474107e-03
## 22 6.155627e-03 0.122962778 6.697240e-01 1.469913892 8.655152e-01
## 23 1.304683e-01 1.032735516 4.379198e-02 1.307285256 6.910733e-02
## 24 1.158130e+01 0.039919437 2.172978e+00 1.687677282 3.293391e+00
## 25 1.771879e-03 1.676697372 4.653051e+00 0.211114075 2.579648e+00
## 26 9.531413e-04 0.002375444 7.572911e-03 1.806078657 5.490705e-01
## 27 6.657853e-03 2.780861152 1.319836e-01 2.497230662 6.139695e-01
## 28 1.471538e-01 0.058132481 4.906989e-01 0.947551461 1.614456e+00
## 29 2.163921e+00 0.179037853 1.068138e+00 0.361629203 2.124898e-01
## 30 1.128279e-01 0.054124427 8.576445e-01 0.104835647 5.885687e+00
## 31 4.487711e-02 9.522094593 3.577836e+00 0.071099145 2.242533e+00
## 32 1.868447e+00 0.162454759 1.371181e+00 0.115772284 3.854117e+00
## 33 8.287217e-01 0.835843825 3.004203e-01 2.304711852 1.222873e+00
## 34 1.714196e+00 0.029707350 8.739611e-02 0.108585261 1.045725e-03
## 35 3.291638e-01 0.193045091 2.101402e-05 0.001320251 1.255433e+01
## 36 2.554878e-01 0.056772433 1.217932e+00 2.701785983 2.260770e-01
## 37 1.117412e+00 0.394265549 1.978928e-03 0.031481062 1.117713e+00
## 38 4.408039e-01 0.976031212 1.006287e-01 0.407427648 9.162580e-02
## 39 2.569609e-01 0.267583631 2.957465e+00 0.141437701 4.173848e-01
## 40 8.750548e-02 0.843899528 1.565918e+00 0.640839663 1.926427e-02
## 41 3.710064e-01 1.084193866 8.140230e+00 0.001815277 2.124093e-01
## 42 1.410683e-01 1.447459658 5.239091e-01 0.865955803 1.336166e-02
## 43 1.426670e+00 0.459416740 1.788929e-01 0.564929703 2.022809e-01
## 44 9.293018e-05 1.955160340 1.418489e+00 1.531681751 3.963719e-01
## 45 8.412233e-03 0.094863402 6.429084e-01 2.842301338 8.986768e-01
## 46 1.519086e+00 0.635184377 1.344324e-01 1.794170070 3.468048e-01
## 47 8.775777e-02 1.164496941 9.835942e-01 0.822788420 3.641776e-03
## 48 1.567454e-02 1.046598962 4.497940e-01 3.405560344 7.640158e-01
## 49 4.027318e-01 0.421682147 3.817350e-01 1.397871874 1.762485e-03
## 50 3.240638e-01 1.290835478 3.973141e-01 2.317351748 2.534985e+00
## 51 6.944866e-01 0.183988706 2.509369e+00 0.005507304 1.364667e+00

```
## 52 8.556139e-01 2.238094955 2.528329e-01 0.162457265 6.409350e-01
## 53 1.968579e-02 0.104690791 1.465974e-02 0.002943842 2.314478e+00
## 54 2.514738e-01 0.043686331 9.480505e-01 3.253284640 3.914904e-02
## 55 9.342293e-01 1.557039176 3.214674e+00 0.373586783 8.903859e-01
## 56 7.587704e-02 0.246169693 1.146947e+00 7.622320520 2.246645e-03
## 57 1.719225e+00 2.986300406 2.821656e-02 0.477962443 6.551068e-03
## 58 9.143266e-01 0.797940770 6.084993e+00 3.689545840 1.733621e-02
## 59 2.616535e+00 2.878987606 2.552584e+00 2.292045944 5.644079e-01
## 60 1.348052e+00 1.828065760 8.475192e-01 0.034453141 1.458370e-01
## 61 4.957093e+00 8.382133994 1.620358e+00 0.588721254 3.370918e+00
## 62 7.609135e+00 0.440318241 7.075794e-02 0.324564861 3.027440e-03
## 63 1.862361e+00 0.173106898 9.651206e-03 0.458210755 2.842130e+00
## 64 4.061246e-01 0.003761680 1.275690e+00 3.093390847 1.123638e+01
## 65 7.540733e-02 0.247357713 1.176137e+00 0.428467608 3.187895e-02
## 66 7.295714e-01 3.680151623 1.748993e-01 0.387032995 1.446618e+00
## 67 9.174333e-01 1.055473824 4.391646e-01 2.019446982 8.916731e-01
## 68 7.127626e+00 0.402131383 9.375776e-02 0.550997219 1.523099e-03
## 69 1.475220e+00 2.451375565 4.570576e-01 0.012723713 2.477621e-01
## 70 3.740715e-01 0.043128127 6.946237e-01 0.761160911 1.150437e+00
## 71 2.077891e-01 0.067442413 3.154464e-02 2.327230181 1.143398e-03
## 72 4.320258e-01 0.165149226 1.714139e+00 2.014738349 2.930054e-04
## 73 5.091340e-01 1.411497826 5.470455e-01 0.115505423 1.077318e-01
## 74 1.038210e+00 0.539134439 1.307064e-01 0.049144191 4.319336e+00
## 75 5.747244e-02 0.639053234 6.726907e-01 0.108069684 5.012421e-03
## 76 5.238457e-01 0.059469130 1.487232e+00 2.760226264 1.229518e-02
## 77 8.827394e-02 0.006250151 1.105331e-01 0.276396713 2.066126e+00
## 78 2.927277e-04 0.086467311 1.287662e+00 0.001900641 2.328136e-02
## 79 1.578904e+00 0.050151720 7.795712e-01 1.266867399 3.133474e-01
## 80 4.931955e-01 0.376868977 5.432737e-02 0.107710317 3.757449e+00
## 81 7.508522e-01 1.453248663 8.824488e-02 1.878045903 9.808463e-01
```

```
mean(res.pca$ind$contrib)
```

```
## [1] 1.234568
```

*#4.4.2. CALIDAD DE REPRESENTACIÓN DE VARIABLES*

```
res.pca$ind$cos2
```

```
##          Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
## 1  7.741287e-02 0.3458198329 1.394115e-01 0.4230022784 1.435348e-02
## 2  4.044184e-02 0.6442184523 6.174930e-02 0.2524671143 1.123295e-03
## 3  5.749883e-02 0.6529936829 9.953718e-02 0.1868169703 3.153343e-03
## 4  2.054715e-02 0.3694743919 2.923259e-01 0.0163863673 3.012662e-01
## 5  8.699505e-01 0.0807426290 1.055640e-02 0.0250477760 1.370271e-02
```

```
## 6  2.651643e-01 0.4402967151 1.168149e-01 0.0795246921 9.819947e-02
## 7  1.218370e-03 0.0632462475 6.350808e-01 0.1047135670 1.957410e-01
## 8  6.300470e-02 0.4393898679 1.267327e-01 0.3566816602 1.419104e-02
## 9  2.838446e-02 0.6803405470 1.903833e-01 0.0994829071 1.408786e-03
## 10 8.923082e-01 0.0313082103 1.443059e-02 0.0541129128 7.840045e-03
## 11 7.124670e-02 0.6975309333 5.327220e-03 0.2257423923 1.527555e-04
## 12 1.913037e-02 0.0022990867 3.383486e-01 0.2025529272 4.376690e-01
## 13 1.571286e-01 0.4070464830 1.665486e-01 0.2679622306 1.314071e-03
## 14 7.390407e-03 0.3191098143 4.173335e-01 0.1928549105 6.331133e-02
## 15 4.094780e-01 0.0091556209 2.013012e-01 0.3675772442 1.248792e-02
## 16 3.166669e-01 0.2056042466 3.759994e-01 0.0002608737 1.014686e-01
## 17 2.745853e-01 0.3004148991 5.212143e-03 0.4156287040 4.158955e-03
## 18 7.953374e-01 0.0002720778 2.028419e-01 0.0001793588 1.369237e-03
## 19 9.706319e-02 0.7455845155 1.171018e-01 0.0070713077 3.317916e-02
## 20 7.000419e-01 0.1962721617 3.084864e-02 0.0512358988 2.160138e-02
## 21 1.755992e-01 0.5821720595 1.745231e-02 0.2232358915 1.540507e-03
## 22 4.020000e-03 0.0614577563 2.539900e-01 0.4519627269 2.285695e-01
## 23 8.206959e-02 0.4971821004 1.599700e-02 0.3871724308 1.757888e-02
## 24 7.720804e-01 0.0020367552 8.412544e-02 0.0529726270 8.878475e-02
## 25 3.454173e-04 0.2501577548 5.267624e-01 0.0193768941 2.033576e-01
## 26 8.829110e-04 0.0016840447 4.073703e-03 0.7876861757 2.056732e-01
## 27 1.831288e-03 0.5853978545 2.108189e-02 0.3233986259 6.829034e-02
## 28 9.339624e-02 0.0282374928 1.808590e-01 0.2831507850 4.143565e-01
## 29 6.810301e-01 0.0431240011 1.952177e-01 0.0535852638 2.704288e-02
## 30 3.661203e-02 0.0134415657 1.616149e-01 0.0160167064 7.723148e-01
## 31 4.335759e-03 0.7040799333 2.007371e-01 0.0032341603 8.761310e-02
## 32 4.244457e-01 0.0282437897 1.808852e-01 0.0123823428 3.540430e-01
## 33 2.571678e-01 0.1985099370 5.413824e-02 0.3367297121 1.534543e-01
## 34 9.320173e-01 0.0123616581 2.759450e-02 0.0277965812 2.299170e-04
## 35 5.926325e-02 0.0265999868 2.197098e-06 0.0001119146 9.140226e-01
## 36 1.078145e-01 0.0183355500 2.984675e-01 0.5368031680 3.857922e-02
## 37 5.921293e-01 0.1598974178 6.089763e-04 0.0078543398 2.395100e-01
## 38 2.988287e-01 0.5063955499 3.961555e-02 0.1300422160 2.511797e-02
## 39 1.064203e-01 0.0848137724 7.112857e-01 0.0275790762 6.990111e-02
## 40 4.482312e-02 0.3308314939 4.658036e-01 0.1545514826 3.990334e-03
## 41 6.168310e-02 0.1379561042 7.859380e-01 0.0001420969 1.428066e-02
## 42 7.174618e-02 0.5634108211 1.547363e-01 0.2073586385 2.748020e-03
## 43 6.397788e-01 0.1576748734 4.658718e-02 0.1192772933 3.668185e-02
## 44 2.902601e-05 0.4673714829 2.572903e-01 0.2252454667 5.006375e-02
## 45 3.901796e-03 0.0336745080 1.731687e-01 0.6206980416 1.685570e-01
## 46 4.950976e-01 0.1584373143 2.544363e-02 0.2753144220 4.570708e-02
## 47 4.525845e-02 0.4596229763 2.945758e-01 0.1997833263 7.594817e-04
## 48 5.241901e-03 0.2678696591 8.735235e-02 0.5362156631 1.033204e-01
```

```
## 49 2.507668e-01 0.2009502743 1.380330e-01 0.4098061257 4.437817e-04
## 50 8.856943e-02 0.2700060845 6.306005e-02 0.2981963396 2.801681e-01
## 51 2.439382e-01 0.0494602579 5.118556e-01 0.0009107768 1.938351e-01
## 52 2.804379e-01 0.5614184217 4.812374e-02 0.0250700372 8.494988e-02
## 53 1.882659e-02 0.0766260970 8.141645e-03 0.0013255324 8.950801e-01
## 54 1.055278e-01 0.0140303642 2.310322e-01 0.6427663323 6.643319e-03
## 55 2.062930e-01 0.2631356875 4.122255e-01 0.0388400166 7.950586e-02
## 56 1.678696e-02 0.0416817008 1.473573e-01 0.7939730195 2.009953e-04
## 57 4.046375e-01 0.5379180181 3.856598e-03 0.0529644005 6.234984e-04
## 58 1.344040e-01 0.0897698809 5.194428e-01 0.2553527824 1.030516e-03
## 59 3.438451e-01 0.2895512440 1.947975e-01 0.1418131974 2.999296e-02
## 60 4.067143e-01 0.4221082914 1.484907e-01 0.0048940521 1.779264e-02
## 61 3.552583e-01 0.4597492434 6.743651e-02 0.0198647832 9.769115e-02
## 62 9.346395e-01 0.0413927963 5.047205e-03 0.0187701299 1.503745e-04
## 63 5.533705e-01 0.0393654993 1.665331e-03 0.0641024101 3.414962e-01
## 64 5.680043e-02 0.0004026460 1.036106e-01 0.2036967194 6.354896e-01
## 65 6.487503e-02 0.1628691043 5.876095e-01 0.1735557219 1.109067e-02
## 66 1.652528e-01 0.6379638204 2.300573e-02 0.0412748990 1.325027e-01
## 67 2.787173e-01 0.2454065028 7.747893e-02 0.2888542393 1.095431e-01
## 68 9.197086e-01 0.0397120755 7.025536e-03 0.0334743163 7.947369e-05
## 69 3.962517e-01 0.5039337038 7.129387e-02 0.0016091076 2.691158e-02
## 70 2.289234e-01 0.0201997299 2.468608e-01 0.2193154016 2.847006e-01
## 71 1.512406e-01 0.0375688604 1.333332e-02 0.7975206754 3.365371e-04
## 72 1.726338e-01 0.0505058681 3.977673e-01 0.3790457587 4.734581e-05
## 73 2.539291e-01 0.5387779510 1.584422e-01 0.0271231212 2.172771e-02
## 74 3.149409e-01 0.1251671466 2.302543e-02 0.0070189663 5.298475e-01
## 75 5.804633e-02 0.4939709575 3.945459e-01 0.0513896844 2.047164e-03
## 76 1.913541e-01 0.0166255171 3.154860e-01 0.4747182771 1.816181e-03
## 77 7.861385e-02 0.0042599677 5.716441e-02 0.1158927730 7.440690e-01
## 78 3.550154e-04 0.0802572977 9.068846e-01 0.0010852766 1.141778e-02
## 79 5.652711e-01 0.0137415636 1.620779e-01 0.2135448827 4.536459e-02
## 80 2.069360e-01 0.1210198173 1.323739e-02 0.0212780144 6.375288e-01
## 81 2.349957e-01 0.3480924549 1.603843e-02 0.2767378555 1.241356e-01
```

```
mean(res.pca$ind$cos2)
```

```
## [1] 0.2
```

***##Cos2 se utiliza como medida de calidad en la representación de los individuos, l os cuales en este caso de aplicación son los cantones de Costa Rica. Asumiendo qu e los individuos cuyo coseno cuadrado sea mayor a 0.50 son representativos, se co ncluiría que: 1) a grandes rasgos, los cantones están bien representados (con la exc epción del puñado que está cerca del origen), 2) Con dos dimensiones se explica el 5 4.8% de la variabilidad del conjunto de datos. Se habla aquí de individuos y no de***

***variables porque los individuos (los cantones) son, en este caso concreto, las unidades de análisis de interés.***

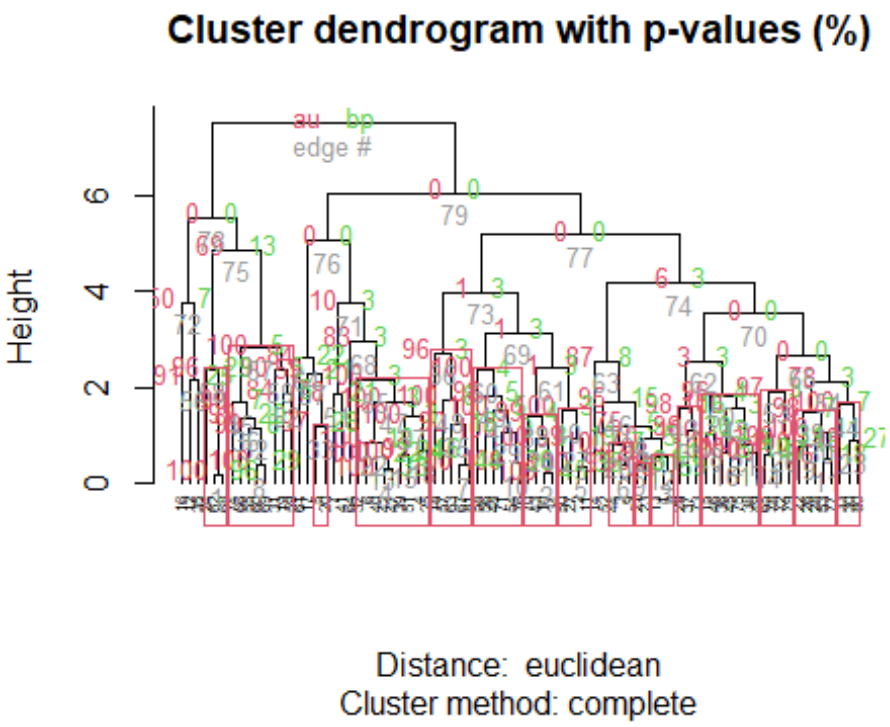
*#4.6. AGRUPAMIENTOS JERÁRQUICOS*

***##Es posible realizar una estimación de clústeres jerárquicos de variables con valor p vía bootstrapping (permutaciones aleatorias aplicadas al diseño muestral del experimento; en caso se apliquen al diseño experimental, se trata de Monte Carlo). Esta estimación permite relacionar las variables de forma jerárquica y detectar visualmente (i.e., de forma exploratoria), aunque con cierta robustez proporcionada por el bootstrapping y los valores p construidos a través de tal remuestreo indicios de relaciones funcionales entre ellas y un valor p asociado a tales indicios.***

```
library(pvclust)
sDATAtrans<-t(sDATA)
set.seed(123) #garantizando la reproductibilidad del experimento
res.pvclust<-pvclust(sDATAtrans, method.hclust = "complete",
      method.dist = "euclidian", nboot = 501, store = TRUE)
```

```
## Bootstrap (r = 0.4)... Done.
## Bootstrap (r = 0.6)... Done.
## Bootstrap (r = 0.8)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.2)... Done.
## Bootstrap (r = 1.4)... Done.
```

```
plot(res.pvclust, hang = -1, cex = 0.5)
pvrect(res.pvclust) #los valores en el dendrograma son valores p de AU (rojo, izquierda), valores de BP (verde, derecha) y etiquetas de grupo (gris, abajo). Los clústeres con AU >= 95 % se indican mediante rectángulos y se considera que están fuertemente respaldados por los datos; sin embargo, esta consideración es provisional porque aquí los resultados no se están analizando a la luz un nivel de confianza (que denota la fiabilidad sobre el valor p obtenido) preestablecido. Ello se lleva a cabo con la sintaxis "pvpick", como se mostrará a continuación.
```




V.

```
hclusters <- pvpick(res.pvclust, alpha=0.95, pv="au", type="geq", max.only=TRUE)
hclusters #la sintaxis "pvpick" encuentra clústeres con valores p relativamente altos/bajos, según los criterios definidos en la sección de conceptualización (seccion 3). Se usan 501 iter aciones, puesto que el número de iteraciones recomendadas para esta clase de métodos son, s egún lo presentado por (Robert & Casella, 2010, págs. 112, 135, 143, 147), 500 o más. Es to se hace para garantizar aproximación asintótica (y tener cierta garantía del cumplimient o del teorema central del límite). Por cuestiones de costo computacional, se escoge 501; esto mismo se puede verificar también en la p. 8 del manual de la librería "factoextra"[9].

## $clusters
## $clusters[[1]]
## NULL
##
## $clusters[[2]]
## NULL
##
## $clusters[[3]]
## NULL
##
## $clusters[[4]]
```

[9] Véase https://cran.r-project.org/web/packages/factoextra/factoextra.pdf.

```
## NULL
##
## $clusters[[5]]
## NULL
##
## $clusters[[6]]
## NULL
##
## $clusters[[7]]
## NULL
##
## $clusters[[8]]
## NULL
##
## $clusters[[9]]
## NULL
##
## $clusters[[10]]
## NULL
##
## $clusters[[11]]
## NULL
##
## $clusters[[12]]
## NULL
##
## $clusters[[13]]
## NULL
##
## $clusters[[14]]
## NULL
##
## $clusters[[15]]
## NULL
##
## $clusters[[16]]
## NULL
##
##
## $edges
##  [1]  9 12 20 33 38 39 40 43 44 50 53 55 58 60 66 67
```

```
res.pvclust$edges$au #número de cada uno de los clústeres generados (que en el dendrog
rama aparecen con letra gris) y su valor p aproximadamente insesgado (au=valor p * 100)

##  [1] 0.999885902 0.999344128 0.999705421 0.997092083 0.999613835 0.992467547
##  [7] 0.996576880 0.999969859 0.964697655 0.999755138 0.955163069 0.984250765
## [13] 0.999714774 0.998055045 0.999189876 1.000000000 0.999913638 0.985145067
## [19] 0.996105679 0.995055676 0.999883811 0.999993828 0.978325512 1.000000000
## [25] 0.698765307 0.997247774 0.994678420 0.998974438 0.998320158 0.999990126
## [31] 0.955813257 0.989477630 0.973589580 0.753047959 0.902869992 0.989665142
## [37] 0.999999799 0.994737586 0.976288081 0.999875816 0.999363128 0.999152124
## [43] 0.976840092 0.999996149 0.994812517 0.946471380 0.996743165 0.998450416
## [49] 0.981374907 0.960851755 0.782105414 0.841198746 0.974736592 0.784956668
## [55] 0.998995321 0.907127682 0.590642741 0.956166907 0.900217183 0.999876325
## [61] 0.005696761 0.029343884 0.865891784 0.835473160 0.000000000 0.955012470
## [67] 0.999999992 0.826660119 0.005696761 0.000000000 0.103436701 0.501518520
## [73] 0.005696761 0.062928668 0.687045459 0.000000000 0.000000000 0.000000000
## [79] 0.000000000 1.000000000

res.pvclust$hclust$order #orden en que se graficaron loss clústeres en el dendrograma

##  [1] 16 12 35 24 62 68 74 66 55 60 69 31 19 59 61 17  5 20  7 41 64 32  6 40 42
## [26] 47 39 51  4 25 18 57 63 67 81 58 56 48 71 54 76 70 43 10 34 50 27  2 11 14
## [51] 52 44  3  9 21 23 13  1  8 29 37 73 15 46 36 45 75 38 49 65 78 72 79 22 26
## [76] 28 53 77 33 30 80

res.pvclust$hclust$merge #merge indica cuáles son los individuos (en este caso son los in
dividuos porque se usó la transpuesta de la matriz sDATA; la sintaxis por defecto utiliza la
s columnas de la matriz imputada) combinados en cada clúster (donde el número de clúster
aparece en el dendrograma con color gris). La razón de la alternancia de signo (que algunos
cantones salgan con signo positivo y otros con negativo) obedece a que es la señalización de
en qué lado se localiza respecto a la parte inicial del dendrograma: a la izquierda se represen
ta con signo negativo y a la derecha con signo positivo.

##       [,1] [,2]
##  [1,]  -62  -68
##  [2,]  -10  -34
##  [3,]   -1   -8
##  [4,]  -42  -47
##  [5,]   -2  -11
##  [6,]   -3   -9
##  [7,]  -67  -81
##  [8,]  -60  -69
##  [9,]  -21  -23
## [10,]  -54  -76
## [11,]  -53  -77
```

```
## [12,] -13   3
## [13,] -40   4
## [14,] -65 -78
## [15,] -39 -51
## [16,] -36 -45
## [17,] -38 -49
## [18,]  -4 -25
## [19,] -22 -26
## [20,] -44   6
## [21,] -75  17
## [22,]  -6  13
## [23,] -30 -80
## [24,] -43   2
## [25,]   9  12
## [26,] -72 -79
## [27,] -27   5
## [28,] -15 -46
## [29,] -71  10
## [30,] -28  11
## [31,] -37 -73
## [32,] -55   8
## [33,]  -5 -20
## [34,]  20  25
## [35,] -63   7
## [36,] -66  32
## [37,]  15  18
## [38,] -70  24
## [39,]  19  30
## [40,] -50  27
## [41,] -48  29
## [42,]  16  21
## [43,] -29  31
## [44,] -33  23
## [45,] -74  36
## [46,] -52  34
## [47,]  22  37
## [48,] -57  35
## [49,] -56  41
## [50,]  28  42
## [51,]  -7 -41
## [52,] -19 -59
## [53,]  14  26
## [54,]  39  44
```

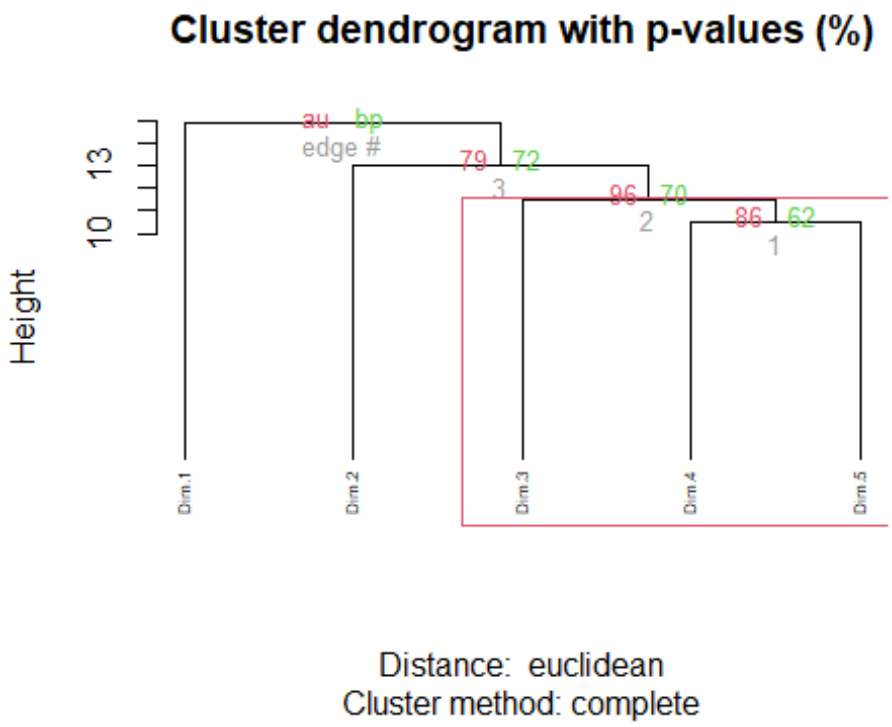
```
## [55,] -32  47
## [56,] -12 -35
## [57,] -17  33
## [58,] -24   1
## [59,] -31  52
## [60,] -58  49
## [61,]  38  40
## [62,]  43  50
## [63,] -14  46
## [64,] -61  57
## [65,]  53  54
## [66,] -18  48
## [67,]  45  59
## [68,] -64  55
## [69,]  60  61
## [70,]  62  65
## [71,]  51  68
## [72,] -16  56
## [73,]  66  69
## [74,]  63  70
## [75,]  58  67
## [76,]  64  71
## [77,]  73  74
## [78,]  72  75
## [79,]  76  77
## [80,]  78  79
```

***##Así, para realizar un análisis con rigor mínimo sobre cuáles cantones se agrupan jerárquicamente en qué número de clústeres y la significancia de estas agrupaciones (dado un nivel de confianza, que en este caso se definió como 0.95), es necesario conocer los números de los clústeres y su valor de probabilidad dado un nivel de confianza preestablecido (esto se conoce a través de la sintaxis "res.pvclust$edges$au"), los pares de individuos agrupados en cada clúster (esto se conoce con la sintaxis "res.pvclust$hclust$merge") y qué clústeres son significativos a determinado nivel de confianza (esto se conoce a través de la sintaxis "pvpick(res.pvclust, alpha=0.95, pv="au", type="geq", max.only=TRUE)").***

***##El procedimiento anterior no utilizó las cargas factoriales, las cuales es posible obtener mediante "ind_coord =res.pca$ind$coord[,a:b]" (donde a y b son números enteros positivos que representan las columnas del dataframe que contiene los resultados del PCA), puesto que, como se verifica en*** (Kassambara, Computing P-value for Hierarchical Clustering, 2021)***, este método permite la independencia del análisis de conglomerados jerárquicos del análisis de componentes principales. Si por determinadas razones esta independencia no se desea, es posible realizar el análisis d***

***e conglomerados jerárquicos utilizando las dimensiones generadas por el embebim iento métrico de tipo PCA, tal como se muestra a continuación.***

```
library(pvclust)
ind_coord =res.pca$ind$coord[,1:5] #esta sintaxis sirve para utilizar las cargas factorial es contenidas en las nuevas dimensiones generadas tras el embebimiento métrico de tipo PC A (que es una aplicación del teorema Johnson-Linderstrauss) y no el grupo de variables orig inales. Se usan las 5 dimensiones generadas porque si se usan menos de 3 es imposible gene rar conglomerados jerárquicos y, en caso usar 3, se genera apenas un conglomerado, por lo cual el valor académico del ejercicio se minimiza.

set.seed(123)
res.pvclust2<-pvclust(ind_coord, method.hclust = "complete",
      method.dist = "euclidian", nboot = 501, store = TRUE)

## Bootstrap (r = 0.49)... Done.
## Bootstrap (r = 0.59)... Done.
## Bootstrap (r = 0.69)... Done.
## Bootstrap (r = 0.79)... Done.
## Bootstrap (r = 0.89)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.1)... Done.
## Bootstrap (r = 1.2)... Done.
## Bootstrap (r = 1.3)... Done.
## Bootstrap (r = 1.4)... Done.

plot(res.pvclust2, hang = -1, cex = 0.5)
pvrect(res.pvclust2)
```




```
hclusters2 <- pvpick(res.pvclust2, alpha=0.95, pv="au", type="geq", max.only=TRU
E)
hclusters2

## $clusters
## $clusters[[1]]
## [1] "Dim.3" "Dim.4" "Dim.5"
##
##
## $edges
## [1] 2

res.pvclust2$edges$au

## [1] 0.8567895 0.9581917 0.7865624 1.0000000

res.pvclust2$hclust$order

## [1] 1 2 3 4 5

res.pvclust2$hclust$merge

##      [,1] [,2]
## [1,]   -4   -5
## [2,]   -3    1
```

```
## [3,]   -2    2
## [4,]   -1    3
```

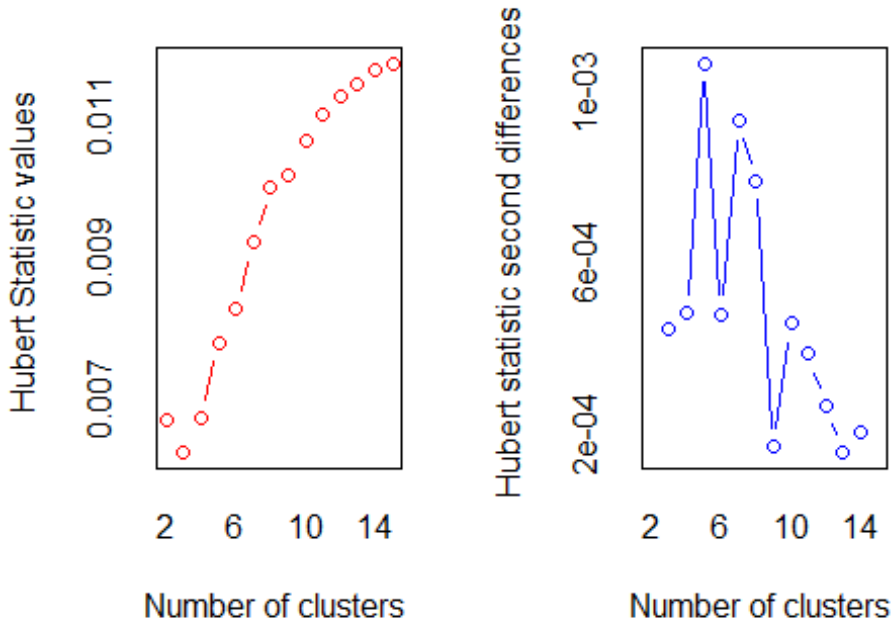
*#4.7. AGRUPAMIENTOS POR K-MEDIAS MEJORADOS VÍA MONTE CARLO*

*#4.7.1. Validación del Número de Clústeres*

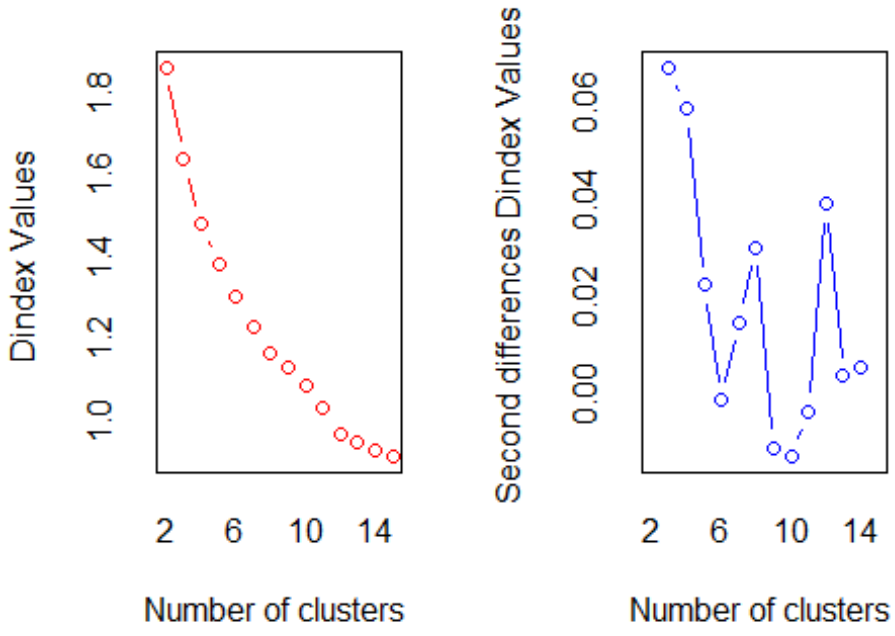
*#4.7.1.1. Validación vía NbClust*

***##Como se señala en https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/, la función NbClust() [en el paquete NbClust R] (Charrad et al. 2014) proporciona 30 índices para determinar el número relevante de clústeres y propone a los usuarios el mejor esquema de agrupamiento a partir de los diferentes resultados obtenidos al variar todas las combinaciones de número de clústeres, medidas de distancia y métodos de agrupación. Puede calcular simultáneamente todos los índices y determinar el número de grupos en una sola llamada de función. Esta sintaxis permite determinar el número de clústeres sin necesidad que el investigador genere ex-ante un análisis de conglomerados por K-medias y seleccione arbitrariamente (o con base en los resultados del PCA) un número óptimo de conglomerados. Así, con esta sintaxis la validación es más robusta porque elimina el sesgo subjetivo (el sesgo estadístico es diferente y es inherente a todo proceso estadístico) y es independiente de los resultados del PCA.***

```
library(NbClust)
NbClust(sDATA, distance="euclidean", min.nc=2, max.nc=15, method="kmeans")
#el número mínimo de clústeres no puede ser inferior a 2 y el número máximo posible es de 15. Se dejó como máximo el número 15 para no condicionar al algoritmo por algún tipo de criterio del investigador.
```




```
## *** : The Hubert index is a graphical method of determining the number of clust
ers.
##             In the plot of Hubert index, we seek a significant knee that correspond
s to a
##             significant increase of the value of the measure i.e the significant peak
in Hubert
##             index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##             In the plot of D index, we seek a significant knee (the significant peak i
n Dindex
##             second differences plot) that corresponds to a significant increase of th
e value of
##             the measure.
##
## *******************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 5 proposed 3 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 4 proposed 12 as the best number of clusters
## * 1 proposed 13 as the best number of clusters
## * 1 proposed 14 as the best number of clusters
## * 1 proposed 15 as the best number of clusters
##
##                 ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
```

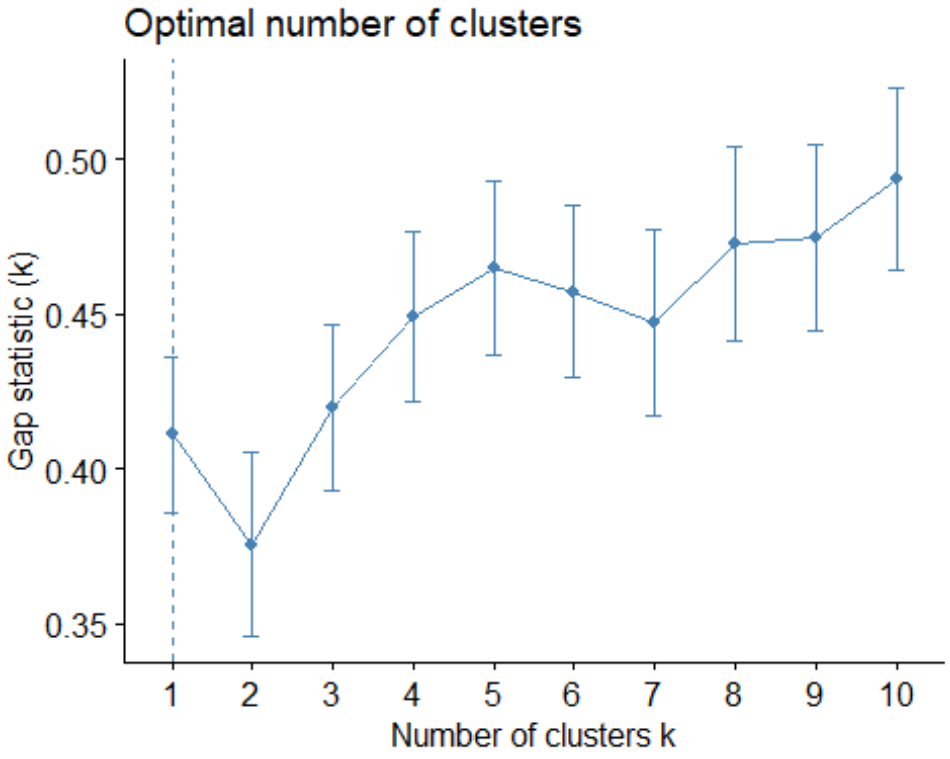
```
##
##
## ************************************************************

## $All.index
##        KL     CH Hartigan     CCC    Scott    Marriot    TrCovW   TraceW
## 2   1.0656 20.6122  17.9503 -1.8232 129.6104 2017305230 4850.2153 317.2302
## 3   0.6276 21.3493  16.0269 -1.7899 307.6798  503739894 3484.6126 258.4950
## 4   1.1585 22.2112  12.9697 -0.9947 291.8931 1088244941 1959.5931 214.4345
## 5   1.4470 22.4119   9.8223 -0.2206 350.1712  828093567 1601.7731 183.5223
## 6   1.8309 21.9189   7.0013  0.1670 368.4503  951558988 1086.0917 162.5183
## 7   0.6525 20.8559   7.5562  0.1606 409.0735  784370818  876.5990 148.6424
## 8   1.5200 20.5005   5.8448  0.5323 461.5368  536061938  738.4071 134.8707
## 9   0.4188 19.8294   8.9693  0.5514 528.0040  298638337  693.4093 124.8727
## 10  1.8119 20.5293   6.0845  1.5494 547.4058  290158143  549.2992 111.0400
## 11  0.6899 20.3771   7.5467  1.8229 569.6375  266821712  518.2010 102.2753
## 12  2.5668 20.9049   4.2399  2.5580 605.7411  203340153  397.7253  92.3220
## 13  2.1324 20.3937   2.9780  2.4567 625.1333  187833473  363.5901  86.9774
## 14 11.7272 19.5861   1.7967  2.0764 637.4325  187153462  314.9796  83.3282
## 15  0.0260 18.5225   4.8183  1.4331 660.1180  162365087  326.3815  81.1520
##    Friedman  Rubin Cindex     DB Silhouette   Duda Pseudot2   Beale Ratkowsky
## 2    3.2653 1.2609 0.4217 1.6006     0.2622 1.8318 -19.9806 -1.3324    0.2595
## 3   15.3881 1.5474 0.3893 1.6656     0.2250 2.3506 -19.5355 -1.6699    0.2846
## 4    7.5046 1.8654 0.3565 1.4536     0.2402 1.6162 -12.2008 -1.1138    0.3267
## 5    8.9719 2.1796 0.3331 1.3383     0.2560 1.3536  -7.3142 -0.7592    0.3215
## 6    8.7900 2.4613 0.3619 1.2612     0.2601 0.9349   0.7663  0.1962    0.3129
## 7   10.3039 2.6910 0.3101 1.2525     0.2553 0.9909   0.1554  0.0266    0.2988
## 8   12.7211 2.9658 0.2958 1.2334     0.2469 2.3760  -9.2659 -1.2084    0.2870
## 9   18.5498 3.2033 0.2737 1.2422     0.2466 1.1244  -1.1065 -0.2597    0.2750
## 10  17.8374 3.6023 0.3832 1.1475     0.2647 1.4061  -2.5992 -0.4519    0.2682
## 11  18.7208 3.9110 0.3716 1.1642     0.2596 1.6098  -5.3033 -1.0539    0.2597
## 12  20.6048 4.3327 0.3653 1.0711     0.2841 0.9825   0.1248  0.0478    0.2529
## 13  21.1699 4.5989 0.4128 1.0439     0.2839 1.7631  -6.0594 -1.2041    0.2452
## 14  21.5967 4.8003 0.4011 1.0738     0.2765 2.2988  -6.2150 -1.5473    0.2377
## 15  23.8578 4.9290 0.4027 1.0676     0.2479 1.6208  -3.0642 -1.0275    0.2303
##       Ball Ptbiserial    Frey McClain   Dunn Hubert SDindex Dindex   SDbw
## 2  158.6151     0.4255  1.0658  0.3699 0.2103 0.0067  2.1850 1.8645 0.8387
## 3   86.1650     0.4241  0.2205  1.2695 0.1902 0.0063  2.1804 1.6397 0.7464
## 4   53.6086     0.4617  0.1129  1.9535 0.1622 0.0067  1.8479 1.4822 0.7417
## 5   36.7045     0.4856  0.3300  2.2827 0.0910 0.0078  1.8096 1.3834 0.6050
## 6   27.0864     0.4785  0.4990  2.6584 0.1201 0.0083  1.6618 1.3070 0.4843
## 7   21.2346     0.4649  0.2490  2.9992 0.1019 0.0092  1.9870 1.2291 0.4987
## 8   16.8588     0.4588  0.1564  3.3756 0.1019 0.0100  2.0383 1.1658 0.4564
```

```
## 9   13.8747     0.4577  0.1504  3.8249 0.1207 0.0102  1.8977 1.1328 0.4140
## 10  11.1040     0.4557  0.3909  4.0942 0.1475 0.0106  1.8993 1.0883 0.3711
## 11   9.2978     0.4346  0.3750  4.7196 0.1333 0.0110  1.9398 1.0307 0.3610
## 12   7.6935     0.4206  0.0837  5.1773 0.1500 0.0113  1.7378 0.9691 0.2760
## 13   6.6906     0.4210  0.2479  5.3124 0.1738 0.0114  1.7174 0.9466 0.2571
## 14   5.9520     0.4115 -1.8350  5.6929 0.1738 0.0117  1.8336 0.9277 0.2505
## 15   5.4101     0.3961  0.0361  6.1446 0.1738 0.0117  2.0461 0.9141 0.2441
##
## $All.CriticalValues
##    CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
## 2          0.4234           59.9295       1.0000
## 3          0.3943           52.2185       1.0000
## 4          0.4095           46.1372       1.0000
## 5          0.3943           43.0035       1.0000
## 6          0.3141           24.0198       0.9624
## 7          0.3943           26.1093       0.9997
## 8         -0.0536         -314.4832       1.0000
## 9          0.0442          216.3177       1.0000
## 10        -0.1969          -54.7084       1.0000
## 11         0.2868           34.8069       1.0000
## 12         0.2177           25.1548       0.9985
## 13         0.2868           34.8069       1.0000
## 14         0.2552           32.1109       1.0000
## 15         0.2177           28.7483       1.0000
##
## $Best.nc
##                     KL      CH Hartigan    CCC    Scott    Marriot   TrCovW
## Number_clusters 14.0000  5.0000  12.0000 12.000   3.0000          3    4.000
## Value_Index     11.7272 22.4119   3.3068  2.558 178.0694 2098070384 1525.019
##                  TraceW Friedman   Rubin Cindex      DB Silhouette   Duda
## Number_clusters  3.0000   3.0000 12.0000 9.0000 13.0000    12.0000 2.0000
## Value_Index     14.6746  12.1228 -0.1554 0.2737  1.0439     0.2841 1.8318
##                PseudoT2   Beale Ratkowsky    Ball PtBiserial   Frey McClain
## Number_clusters   2.0000  2.0000    4.0000  3.0000     5.0000 2.0000  2.0000
## Value_Index     -19.9806 -1.3324    0.3267 72.4501     0.4856 1.0658  0.3699
##                   Dunn Hubert SDindex Dindex    SDbw
## Number_clusters 2.0000      0  6.0000      0 15.0000
## Value_Index     0.2103      0  1.6618      0  0.2441
##
## $Best.partition
##  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 2 2 1 2 1 2 2 2 2 2 1 2
## [39] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 1 1 2 1 2 1 2 1 2 1 1 2 2 2 1 1 2 2
## [77] 2 2 2 2 2
```

***##Se concluye que el número óptimo de clústeres, estimada su similaridad median te la distancia euclidiana y agrupados por el método de K-medias, es 2.***

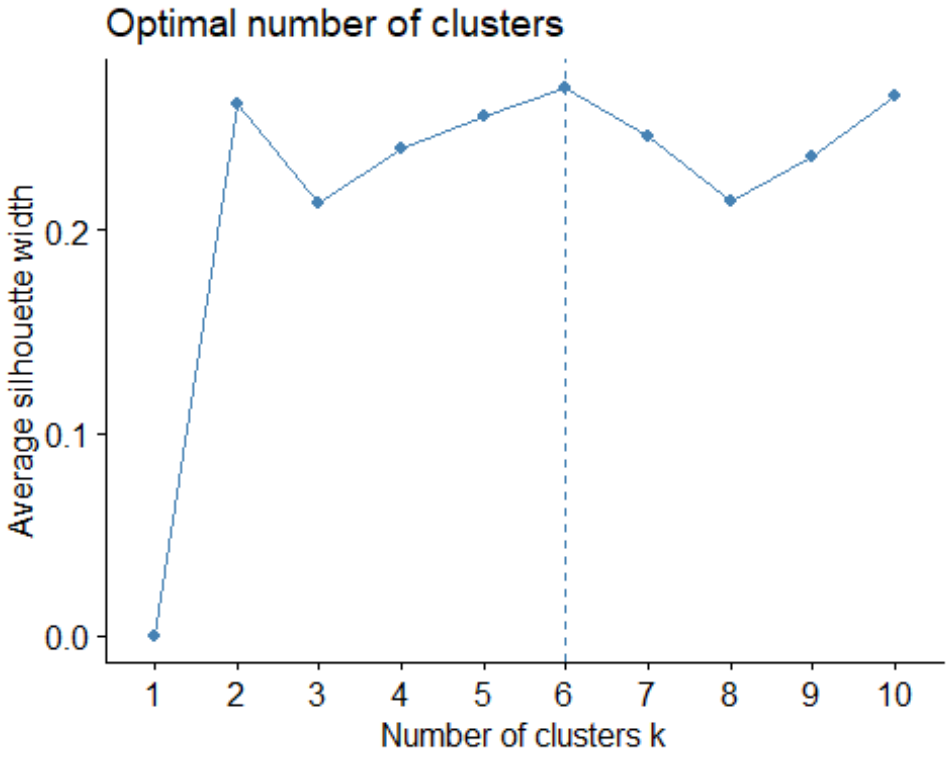
*#4.7.1.2. Validación vía Elbow, Silueta y Estadístico de Brecha*

```
library(factoextra)
fviz_nbclust(ind_coord, kmeans, method=c("gap_stat")) #Método del Estadístico de B
recha
```

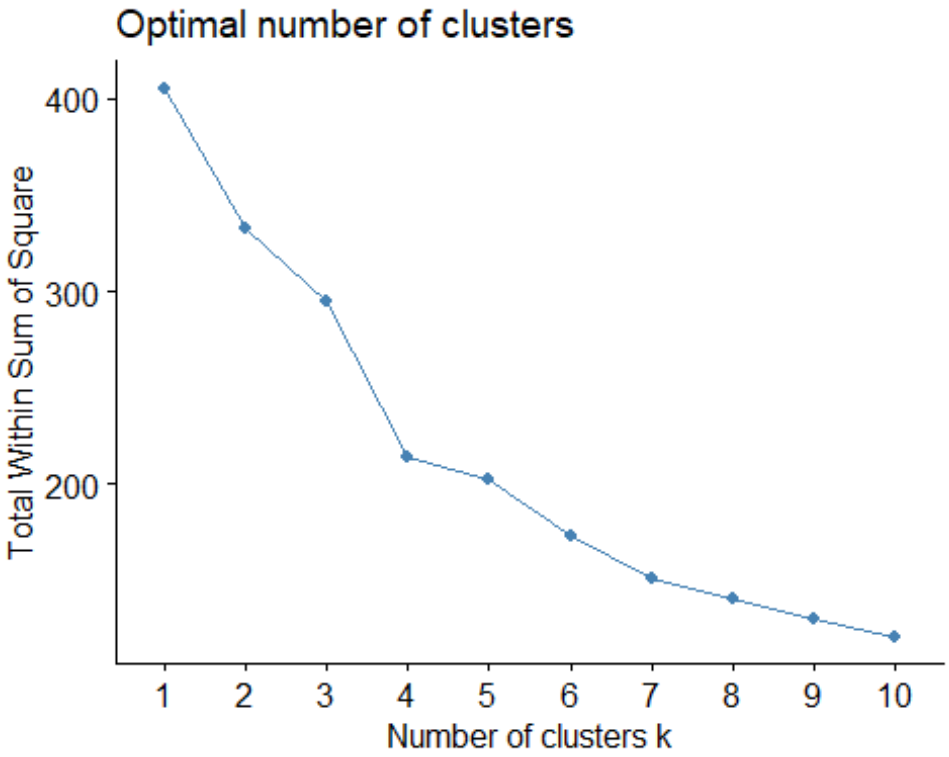



VI.

```
fviz_nbclust(ind_coord, kmeans, method=c("silhouette")) #Método de la Silueta
```

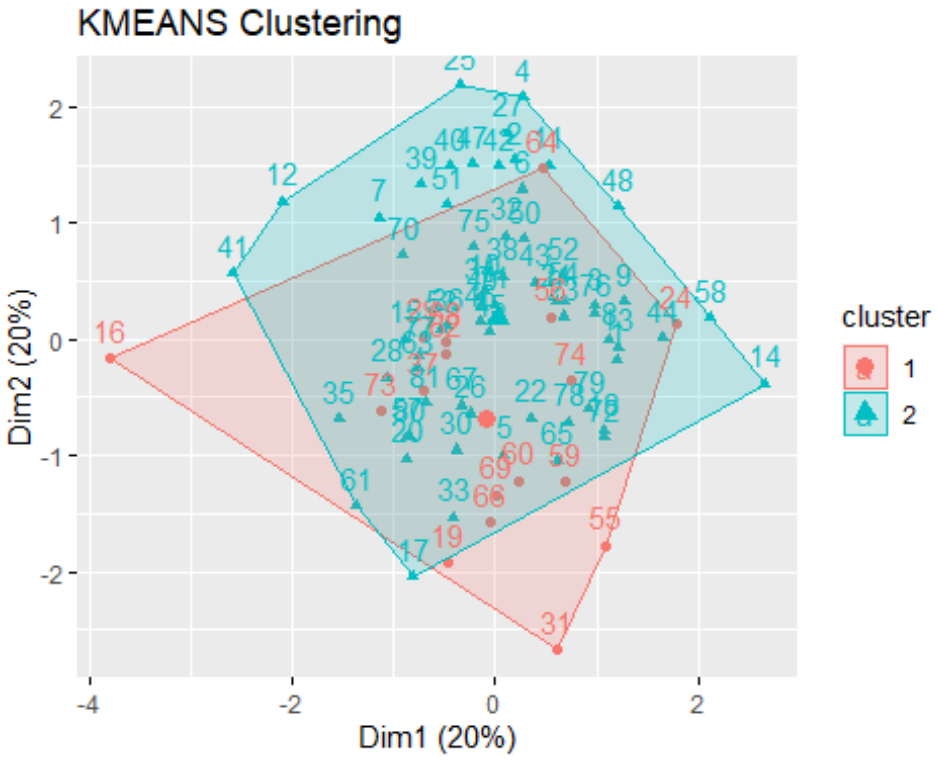



```
fviz_nbclust(ind_coord, kmeans, method=c("wss")) #Método de Elbow
```



***##Como puede observarse, la función de validación NbClust no sólo es más robust a sino que además facilita al investigador la decisión; lo mismo puede afirmarse d el método de la silueta y del estadístico de brecha frente al método del codo (Elbo w), que los primeros no sólo son más robustos sino de más fácil interpretación.***

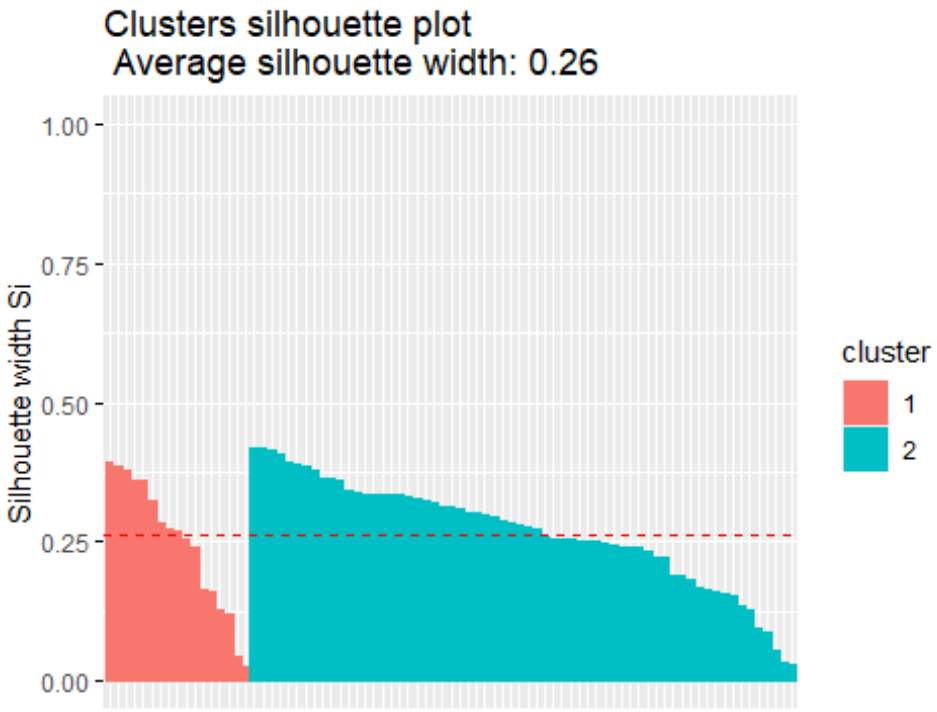
```
library(factoextra)
set.seed(123)
res.km <- eclust(ind_coord,"kmeans", k=2, nboot = 501) #Se seleccionan 2 clústeres, p
uesto que con el método de NbClust se determinó tal magnitud (criterio que coincide con la
selección de dimensiones en el PCA anteriormente hecha con base en, además de criterios p
uramente técnicos, bajo la lógica de la significancia teórica de las dimensiones -que la variab
le tasa de vacunación no se repitiera, dada su alta participación en la dimensión 2 y 3-).
```




***##Teniendo un número de clústeres de partida (mediante la sintaxis "eclust"), o bi en, conociendo con antelación el número óptimo de clústeres (mediante la sintaxis "NbClust"), es posible estudiar los resultados del método silueta con mayor detall e mediante las sintaxis que se presentan y explican a continuación.***

```
fviz_silhouette(res.km) #La gráfica resultante proporciona el coeficiente silueta global de
los conglomerados
```

```
##   cluster size ave.sil.width
## 1       1   17         0.24
## 2       2   64         0.27
```




```
res.km$silinfo$clus.avg.widths #coeficiente silueta para cada clúster (anchura promedio
)

## [1] 0.2447952 0.2667615

res.km$silinfo$widths #Localización de la observación en uno de los 2 clústeres, vecino
más cercano (a qué clúster tiene esa observación más cerca)  y coeficiente silueta individuali
zado (según clúster)

##    cluster neighbor  sil_width
## 69       1        2 0.39400572
## 59       1        2 0.38537410
## 60       1        2 0.37693389
## 62       1        2 0.36033726
## 68       1        2 0.35887536
## 66       1        2 0.32280612
## 19       1        2 0.28414055
## 24       1        2 0.27415765
## 55       1        2 0.26772027
## 74       1        2 0.25595165
## 31       1        2 0.23980866
## 29       1        2 0.16347767
## 16       1        2 0.16071040
## 73       1        2 0.12792715
```

```
## 37      1       2 0.12191541
## 64      1       2 0.04266076
## 56      1       2 0.02471600
## 10      2       1 0.41989864
## 43      2       1 0.41759737
## 34      2       1 0.41378874
## 6       2       1 0.40796425
## 42      2       1 0.39300675
## 11      2       1 0.38945401
## 2       2       1 0.38420906
## 47      2       1 0.37808563
## 40      2       1 0.36447737
## 79      2       1 0.36304214
## 70      2       1 0.35977449
## 51      2       1 0.34036624
## 50      2       1 0.33930863
## 72      2       1 0.33616859
## 63      2       1 0.33599531
## 3       2       1 0.33578239
## 39      2       1 0.33425359
## 9       2       1 0.33321844
## 75      2       1 0.33129924
## 32      2       1 0.32770261
## 23      2       1 0.32231389
## 44      2       1 0.32087011
## 27      2       1 0.31304443
## 18      2       1 0.31271106
## 21      2       1 0.30865277
## 8       2       1 0.30276825
## 53      2       1 0.30237426
## 22      2       1 0.29700579
## 4       2       1 0.29428226
## 5       2       1 0.28797767
## 80      2       1 0.28424547
## 38      2       1 0.28132518
## 26      2       1 0.27791479
## 1       2       1 0.27315473
## 76      2       1 0.25854818
## 14      2       1 0.25465965
## 33      2       1 0.25435715
## 25      2       1 0.25305031
## 45      2       1 0.25194548
## 77      2       1 0.25116890
```

```
## 30      2       1 0.25072684
## 13      2       1 0.24808916
## 49      2       1 0.24310324
## 20      2       1 0.24139634
## 54      2       1 0.24132659
## 52      2       1 0.23981848
## 48      2       1 0.23149153
## 58      2       1 0.22374173
## 71      2       1 0.22197657
## 36      2       1 0.18783438
## 67      2       1 0.18783380
## 28      2       1 0.18265587
## 17      2       1 0.16876800
## 81      2       1 0.16403275
## 41      2       1 0.16072763
## 57      2       1 0.15710458
## 12      2       1 0.15330702
## 46      2       1 0.13377740
## 78      2       1 0.12923711
## 61      2       1 0.09352273
## 35      2       1 0.08721727
## 7       2       1 0.05383683
## 15      2       1 0.03406294
## 65      2       1 0.02938128
```

```
res.km$silinfo$avg.width #coeficiente silueta promedio del total de clústeres (2)
```

```
## [1] 0.2621513
```

```
res.km$centers #los promedios alrededor de los cuales se generan los conglomerados, embe
bidos en el nuevo espacio tridimensional creado
```

```
##       Dim.1      Dim.2       Dim.3       Dim.4      Dim.5
## 1 -1.5134286  1.1402781 -0.22803755 -0.33459527 -0.3746948
## 2  0.4020045 -0.3028864  0.06057248  0.08887687  0.0995283
```

```
res.km$cluster #localización de cada cantón dentro de cada uno de los dos clústeres genera
dos y previamente validados
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  1  2  2  1  2  2  2  2  1  2  2
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
##  2  2  1  2  1  2  2  2  2  2  1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
## 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
##  2  2  1  1  2  2  1  1  2  1  2  1  2  1  2  1  1  2  2  2  1  1  2  2  2  2
```

```
## 79 80 81
##  2  2  2
```

```
res.km$size #tamaño de cada clúster (cuántos cantones están localizados en cada clúster)
```

```
## [1] 17 64
```

***###NOTA: Si se utiliza el método tradicional (no mejorado por boostrapping, y es te mejoramiento permite tener una mayor fiabilidad respecto a los resultados obte nidos -no necesariamente para este caso, pero sí en general-) se podrían obtener un poco diferentes, pero, en tal caso, son más confiables estos resultados.***

## V. CONCLUSIONES Y RECOMENDACIONES

### III.I. Análisis de Componentes Principales

Se concluye que debe trabajarse únicamente con las dimensiones 1 y 2, siguiendo los siguientes criterios: el criterio del autovalor mayor a la unidad y el criterio de que la varianza explicada debe ser mayor al 8-9% (puede variar según el caso de aplicación); respecto al gráfico de sedimentación, debe decirse que aunque esta gráfica sirve para seleccionar el número de componentes que se usarán con base en la magnitud (el "tamaño") de los valores característicos y, puesto que el patrón ideal a buscar para tales fines es una curva pronunciada, seguida de una inflexión y luego de una línea recta (en donde se deben utilizar los componentes en la curva pronunciada antes del primer punto que inicia la tendencia de línea), no procede utilizar este criterio puesto que no parece existir tal punto de inflexión.

Las conclusiones sobre la formación de las dimensiones se presentan a continuación.

- DIMENSIÓN 1: La variable que mayor aporta a la conformación de la primera dimensión es la de mortalidad infantil (34.242%), seguida por el acceso a redes de cuido de personas de tercera edad (30.032%), luego por desnutrición infantil (15.961%), luego por sobrepeso y obesidad infantil (14.377%) y finalmente por tasa de vacunación (5.389%), con un cos2 en todos los casos (salvo para la tasa de vacunación) mayor a 0.10 y en el caso

de las dos variables que más contribuyen en esta dimensión tienen un cos2 mayor a 0.45; siendo rigurosos, i.e., requiriendo un cos2 mayor o igual a 0.5, únicamente la mortalidad infantil podría tomarse como representativa.

- DIMENSIÓN 2: Para esta dimensión, la variable sobrepeso y obesidad infantil es quien realiza la mayor contribución (38.586%), seguida de la desnutrición infantil (30.625%), luego por la tasa de vacunación (23.423%), luego por la mortalidad infantil (6.493%) y finalmente por el acceso a redes de cuido (que es despreciable, apenas contribuye en un 0.873%).
- DIMENSIÓN 3: Para esta dimensión, salvo la tasa de vacunación (que tienen una contribución de aproximadamente el 70% y un cos2 de 0.624), las demás variables no pueden considerarse ni bien representadas ni que tengan una contribución realmente importante, por lo que puede decirse, grosso modo, que la dimensión 3 es tendencialmente equivalente a la tasa de vacunación. Por ello, es natural que sea excluida del análisis.

Respecto al análisis de individuos (cantones), los cantones que más contribuyen son Tarrazú (7.300232e+00%), Turrubares (4.889481e+00%), Dota (2.298290e+00%), Curridabat (6.371242e+00%), León Cortés (6.839104e+00%), Orotina (2.163921e+00%), Abangares (2.616535e+00%), Tilarán (4.957093e+00%) y Nandayure (7.609135e+00%). Es natural que individualmente estudiados, los cantones tengan cifras bajas en comparación al que tenían las variables en el análisis anterior, lo cual obedece a que son 81 cantones (en contraste con las 5 variables), por lo que es esperable que esto ocurra (es esperable que el 100% se tenga que distribuir entre más elementos en consideración -81 cantones versus 5 variables- y que, por tanto, la contribución de cada individuo sea menor en comparación a la de las variables). Puesto que cos2 (coseno cuadrado) se utiliza como medida de calidad en la representación de los individuos. Asumiendo que los individuos cuyo coseno cuadrado sea mayor a 0.10 son representativos, se concluiría que todos los individuos alrededor del origen (por ejemplo, cantones

como Naranjo, San Ramón o Acosta) estarían mal representados por el modelo generado vía PCA, que es una proporción considerable del total de cantones, tal como puede apreciarse a nivel de la gráfica antes presentada.

### III.II. Análisis de Conglomerados Jerárquicos

Con relación al análisis jerárquico de los componentes principales, los conglomerados verdaderamente significativos (que poseen un valor p elevado dado un nivel de confianza aquí definido de 0.95) según los resultados de la función 'pvclust' son los que vinculan al componente principal 1 (aquí definido como "Atención médica para preservación de la vida") y al 2 (aquí definido como "Calidad de alimentación e inoculación de niños"), así como al 3 y al 4. Sin embargo, por los motivos expuestos en la subsección anterior, estas dos dimensiones no son estadísticamente significativas.

Adicionalmente, desde el método del estadístico de brecha se concluye que la cantidad óptima de clústeres es 1, sin embargo, esto carece de sentido en términos de los fines de esta investigación y el número inmediato de dimensiones óptimas es 2, lo cual es congruente con los resultados de la metodología proporcionada por la librería 'NbClust', puesto que la regla de la mayoría indica que el mejor número de conglomerados es 2. Lo sostenido en ese párrafo aplica también para el número óptimo de conglomerados por K-Medias.

### III.III. Análisis de Conglomerados por K-Medias

Se explica un 40% de la variabilidad con las dos dimensiones utilizadas al realizar dentro de ellas un análisis de conglomerados jerárquicos por K-medias. El clúster 1 contiene 17 observaciones y un coeficiente silueta promedio de 0.24, mientras que el clúster 2 contiene 64 observaciones y un coeficiente silueta promedio de 0.27. El coeficiente silueta global es de 0.26. Lo anterior significa que los datos, aunque no están perfectamente separados (perfectamente separados es un coeficiente igual 1 y mal separados es inferior a cero, aunque evidentemente es deseable que se acerque

a 1), tienen la suficiente separación para considerar una primera aproximación válida al método descriptivo utilizado, es decir, las distancias entre clústeres son estadísticamente significativas. El valor promedio (el centro del clúster) de la nueva variable "Atención médica para preservación de la vida" (DIM1) en el primer conglomerado (que contiene a los cantones especificados en la salida del modelo mediante la sintaxis "res.km$cluster") es de -1.51, mientras que en el segundo conglomerado su valor promedio es de 1.14. Por su parte, el valor promedio (el centro del clúster) de la nueva variable "Calidad de alimentación e inoculación de niños" (DIM2) en el primer conglomerado es de 0.40, mientras que en el segundo conglomerado su valor promedio es de -0.30. Puesto que la DIM1 está conformada por "Mortalidad Infantil" y "Acceso a redes de cuido de personas de tercera edad", no es posible concluir respecto a si es deseable que el valor promedio de DIM1 sea positivo o sea negativo y, por consiguiente, tampoco es posible concluir sobre el valor promedio en cuestión. Por otro lado, puesto que DIM2 está conformada por "Tasa de vacunación" y por "Sobrepeso y obesidad infantil", ocurre lo mismo con relación a determinar la deseabilidad antes mencionada. Esto muestra que lo recomendable es trabajar con variables cuyo sentido de variación tenga una interpretación equivalente, con el fin que puedan extraerse conclusiones (aunque sea provisionales) con implicaciones prácticas claras.

### III.IV. Elección del Mejor Método Descriptivo

Se escoge como mejor método descriptivo el análisis de conglomerados por K-medias. Esto se justifica por las siguientes razones:

1. Permite vincular explícitamente a los cantones individualmente considerados con las nuevas dimensiones construidas vía PCA, lo que permite mayor facilidad en el análisis.

2. Se dispone en la biblioteca CRAN de una metodología estadística-computacional más rigurosa para la validación del número óptimo de conglomerados por K-medias que para la de conglomerados jerárquicos [para el caso de los conglomerados jerárquicos esta validación no es tan robusta (la de K-medias usa 30 tipos de validación diferentes) ni metodológicamente tan clara, aunque el resultado es el mismo, salvo la consideración que el segundo conglomerado significativo (el que relaciona la dimensión 3 con la dimensión 4) no es, en realidad (a la luz del análisis PCA realizado), verdaderamente significativo (hecho revelado en la sección en que se hizo PCA al estudiar el objeto "res.pca" mediante diferentes sintaxis)].

## VI. REFERENCIAS


Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. *WIREs Computational Statistics*, 433-459. Retrieved from https://personal.utdallas.edu/~herve/abdi-awPCA2010.pdf

Bock, T. (2021, Octubre 22). *What is Hierarchical Clustering?* Retrieved from Segmentation: https://www.displayr.com/what-is-hierarchical-clustering/

Cross Validated. (2015, Marzo 29). *Loadings vs eigenvectors in PCA: when to use one or another?* Retrieved from StackExchange: https://stats.stackexchange.com/questions/143905/loadings-vs-eigenvectors-in-pca-when-to-use-one-or-another

Cross Validated. (2020, Agosto 28). *How are eigenvalues/singular values related to variance (SVD/PCA)?* Retrieved from StackExchange: https://stats.stackexchange.com/questions/479485/how-are-eigenvalues-singular-values-related-to-variance-svd-pca

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R.* New York: Springer.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2016). *Handbook of Cluster Analysis.* Boca Raton: CRC Press.

INCAE Business School. (2022, Junio 24). *Índice de Progreso Social Cantonal 2019*. Retrieved from Proyectos:

https://www.incae.edu/es/clacds/proyectos/indice-de-progreso-social-cantonal-2019.html

Jollife, I. (2002). *Principal Component Analysis.* New York: Springer-Verlag.

Kassambara, A. (2017, Septiembre 23). *PCA - Principal Component Analysis Essentials*. Retrieved from Statistical Tools for High-throughput Data Analysis: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials

Kassambara, A. (2021, Agosto 1). *Computing P-value for Hierarchical Clustering*. Retrieved from Datanovia: https://web.archive.org/web/20210801010529/https://www.datanovia.com/en/lessons/computing-p-value-for-hierarchical-clustering/

Kassambara, A. (2022, Abril 2). *Determining The Optimal Number Of Clusters: 3 Must Know Methods* . Retrieved from Datanovia: https://web.archive.org/web/20220402102456/https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Kumar, S. (2020, Octubre 18). *Silhouette Method — Better than Elbow Method to find Optimal Clusters*. Retrieved from Towards Data Science: https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

Mathematics. (2012, Octubre 29). *What is the difference between eigenfunctions and eigenvectors of an operator?* Retrieved from StackExchange: https://math.stackexchange.com/questions/223226/what-is-the-difference-between-eigenfunctions-and-eigenvectors-of-an-operator

Nabi, I. (2020). *Sobre los Estimadores de Bayes, el Análisis de Grupos y las Mixturas Gaussianas.* Documento inédito. Retrieved from https://marxianstatistics.files.wordpress.com/2020/12/sobre-los-estimadores-de-bayes-el-analisis-de-grupos-y-las-mixturas-gaussianas-isadore-nabi.pdf

Nabi, I. (2022, Abril 28). *EMBEBIMIENTOS MÉTRICOS: ANÁLISIS DE COMPONENTES PRINCIPALES Y ANÁLISIS DE CONGLOMERADOS. SU INTERPRETACIÓN CONCEPTUAL, INTUICIÓN-LÓGICA GEOMÉTRICA, FORMALISMO MATEMÁTICO Y APLICACIONES EN RSTUDIO Y MINITAB.* Retrieved from Marxist Statistics: https://marxianstatistics.files.wordpress.com/2022/04/embebimientos-metricos.-analisis-de-componentes-principales-y-analisis-de-conglomerados-isadore-nabi.pdf

R CODER. (2022, Julio 5). *Set seed in R*. Retrieved from Introduction to R: https://r-coder.com/set-seed-r/

Ranganathan, M. (2022, Mayo 20). *Operators, Eigenfunctions and the Schrödinger Equation.* Retrieved from Indian Institute of Technology Kanpur: https://home.iitk.ac.in/~madhavr/CHM102/Physical/Lec2.pdf

Robert, C. P., & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. New York: Springer.

stack overflow. (2017, Abril 23). *p-values in pvclust & results in hclust*. Retrieved from Questions: https://stackoverflow.com/questions/43576210/p-values-in-pvclust-results-in-hclust

Universitat de Girona. (2009, Octubre 23). *Número de factores a conservar*. Retrieved from Análisis factorial: https://web.archive.org/web/20091023060436/http://www3.udg.edu/dghha/cat/secciogeografia/prac/models/factorial(5).htm