Universidad Nacional del Oeste

Licenciatura en Informática

Explotación de Datos

Integrantes Grupo 1:

- Robledo Alan
- Farías Gonzalo
- Romano Diego

Docentes:

- Perez Silvia
- Mendoza Dante

30 de septiembre de 2023

Índice

1. Introducción

En este informe, se presenta un análisis de las provincias de Argentina basado en diez indicadores económicos y sociales. Utilizamos el Análisis de Componentes Principales (ACP) para simplificar los datos y revelar patrones, seguido de la aplicación de clustering para agrupar las provincias según su desarrollo. Estos resultados tienen implicaciones significativas para la formulación de políticas públicas, ya que ayudan a identificar áreas de enfoque para la asignación de recursos en infraestructura, educación y programas de bienestar, promoviendo un desarrollo más equitativo en todo el país.

2. Analisis de componentes principales (ACP)

El Análisis de Componentes Principales (ACP) es una técnica estadística utilizada en análisis de datos multivariados para reducir la dimensionalidad de un conjunto de datos, manteniendo la mayor cantidad posible de información importante. Su objetivo principal es simplificar la interpretación de datos complejos al transformar un conjunto de variables correlacionadas en un conjunto más pequeño de variables no correlacionadas llamadas componentes principales. Estos componentes principales se ordenan en función de la cantidad de variabilidad que explican en los datos originales, de manera que el primero explica la mayor parte de la variabilidad y los siguientes explican cada vez menos.

2.1. ACP en R

El ACP realizado en R resultó en una representación más simplificada de nuestros datos originales, lo que nos permitió explorar patrones y relaciones de manera más eficiente. Los resultados de este análisis servirán como base para el posterior proceso de clustering. Lo primero que debemos hacer es verificar la idoneidad de los datos. Esto implica asegurarnos de que no haya valores nulos (lo cual no ocurre en este caso), que las variables estén correlacionadas y que no haya valores atípicos. En este caso, hemos identificado tres valores atípicos, uno de los cuales es especialmente destacado, como se muestra en el gráfico a continuación.

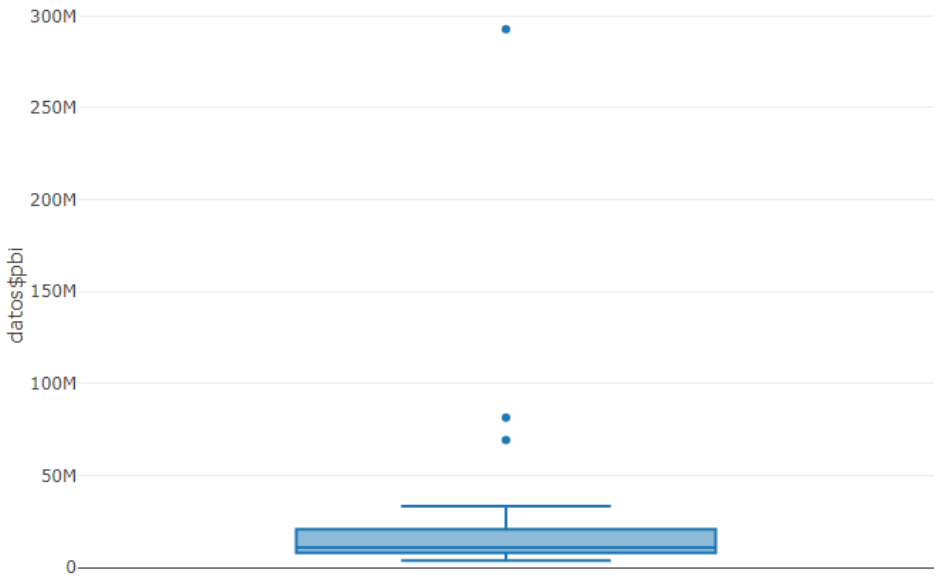

Figura 1: Boxplot

Sin embargo, es importante destacar que este valor atípico está justificado, ya que corresponde a la provincia más importante de Argentina, Buenos Aires.

2.2. ¿Tenemos la correlación entre los componentes necesaria para poder simplificarlos?.

Para poder reducir la dimensionalidad de nuestro conjunto de datos, primero debemos observar y comprobar si existe una correlación en nuestro conjunto de datos, como podemos ver en el siguiente gráfico esto es cierto. Por ejemplo, pbi y población esta correlacionadas muy fuertemente, y a la vez estas se correlacionan, débilmente, con las demás variables. Otro ejemplo es la correlación fuertemente inversa que existe entre las variables cinesporcantidaddehabitante, analfabetismo, pobreza y faltaatenciónmedica.

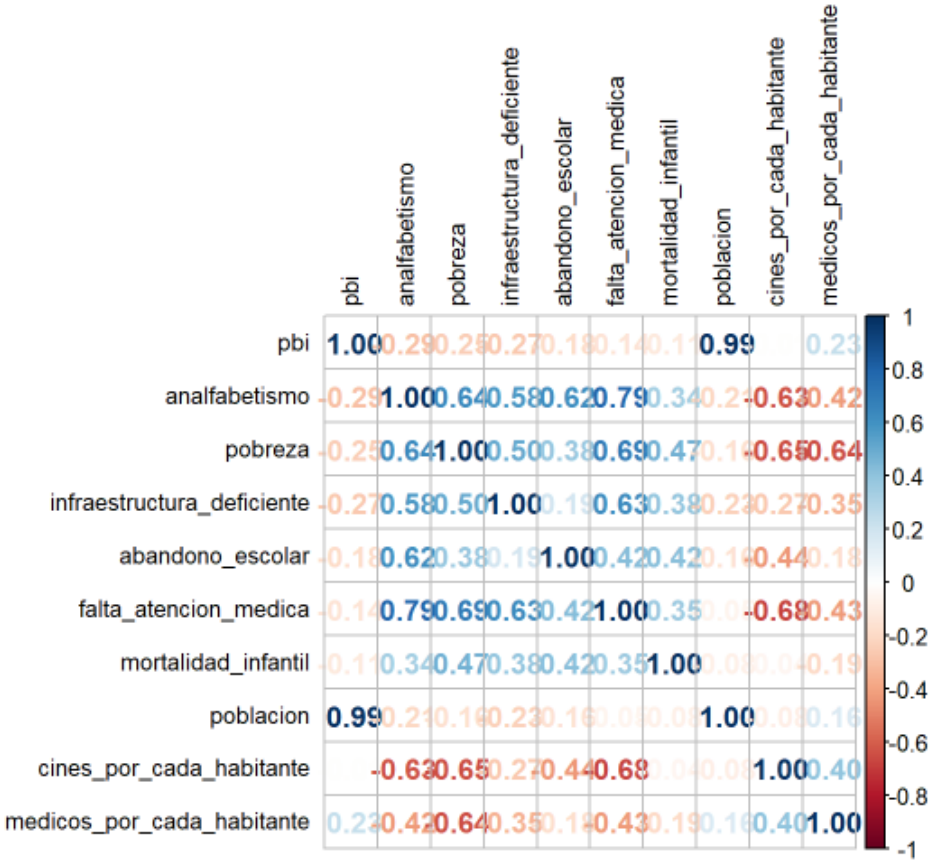

Figura 2: Corrplot

Y así podemos observar más variables que están relacionadas directa o indirectamente. Aunque, para sacarnos la duda, podemos realizar diferentes test que nos comprueben esto, por ejemplo el test de Bartlett:

chisq [1] 174.209

p.value [1] 4.224723e-17

df [1] 45

Como podemos observar el test de Bartlett, el p.value nos da un resultado muy chico, esto nos confirma que es posible reducir la cantidad de variables para poder simplificar la información. No contento con esto, también realizamos el test KMO(Kaiser-Meyer-Olkin)que es otra prueba estadística utilizada para evaluar la idoneidad de aplicar el Análisis de Componentes Principales (PCA)a nuestros datos. El Test KMO mide la adecuación de los datos para el PCA, evaluando si los datos tienen suficiente variabilidad y estructura para justificar la reducción de dimensionalidad mediante PCA. El resultado del Test KMO es un valor que varía entre 0 y 1, donde valores más cercanos a 1 indican una mayor adecuación para el PCA. A continuación podemos ver como también se cumplen con este requisitos:

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(datos))
## Overall MSA =  0.62
## MSA for each item =
##                          pbi                analfabetismo
##                         0.42                         0.72
##                     pobreza infraestructura_deficiente
##                         0.76                         0.58
##            abandono_escolar      falta_atencion_medica
##                         0.52                         0.81
##         mortalidad_infantil                    poblacion
##                         0.50                         0.39
##    cines_por_cada_habitante medicos_por_cada_habitante
##                         0.70                         0.76
```

Figura 3: test kmo

2.3. ¿Qué CP elegir?

Para determinar la cantidad adecuada de componentes principales, primero evaluamos su impacto en la explicación de los datos originales a través de los siguientes gráficos:

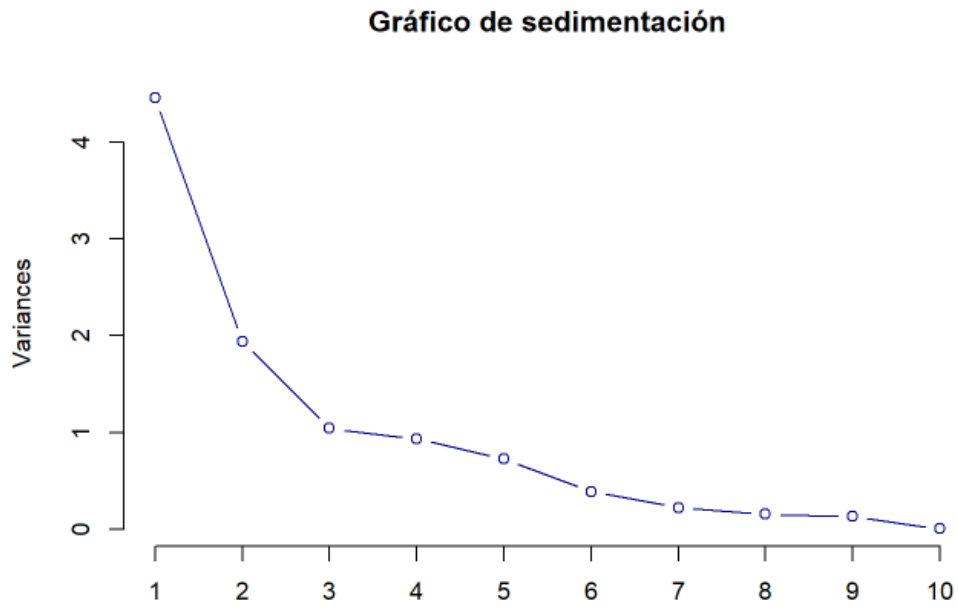

Figura 4

En el primer gráfico, se observa un quiebre en la variabilidad que cada variable aporta a partir de la tercera variable.

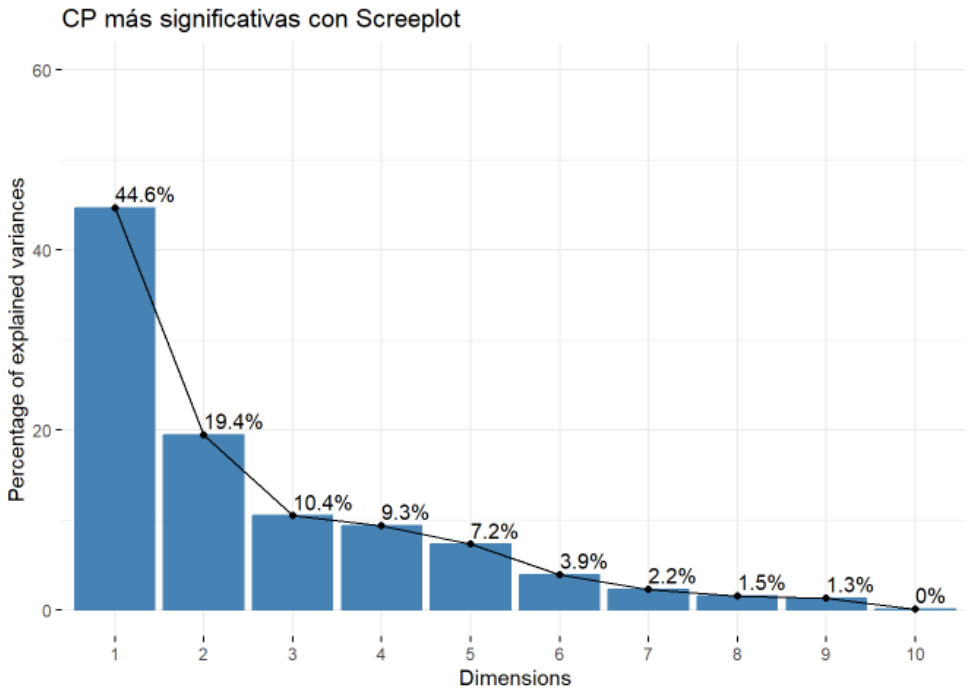

Figura 5

En el segundo gráfico, se muestra el porcentaje de variabilidad que aporta cada variable. A partir de estos datos, podemos concluir que 3 o 4 variables serían suficientes para explicar el 84

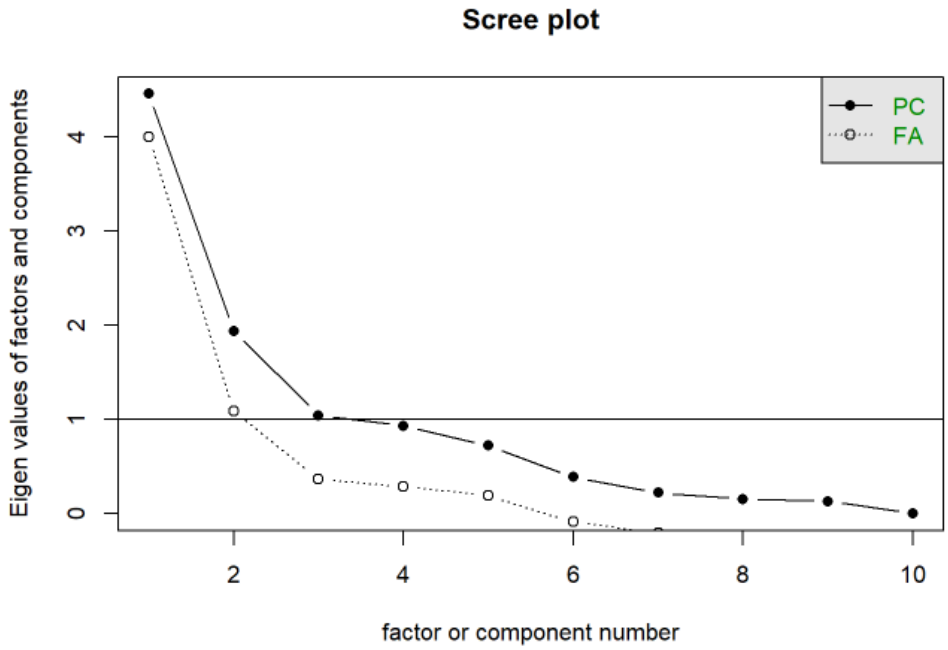

Figura 6

El tercer gráfico indica que elegir 3 CP sería la opción más adecuada, ya que explicaría la mayor cantidad de información con la menor cantidad de componentes necesarios.

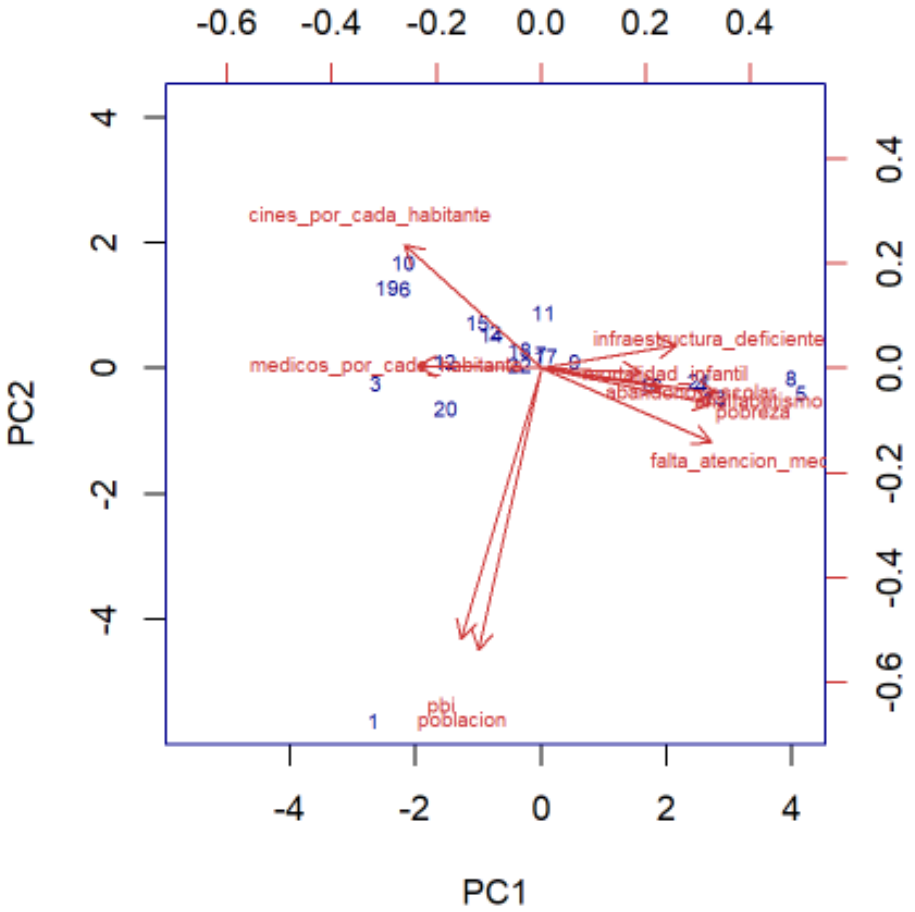

Figura 7: Biplot

En este gráfico, se muestra cómo cada variable se ve afectada por los componentes principales.

3. Clustering

El clustering es una técnica valiosa para explorar patrones y estructuras ocultas en datos, lo que puede llevar a una mejor comprensión de los datos y a la toma de decisiones más informada en diversos campos. Al agrupar objetos o datos similares, el clustering facilita la organización y la interpretación de información compleja, lo que puede ser esencial para la planificación estratégica y la formulación de políticas.

El proceso de clustering en R nos permitió agrupar las provincias de Argentina en segmentos coherentes en función de su desarrollo socioeconómico. Estos clusters proporcionan información valiosa para la formulación de políticas públicas, ya que ayudan a identificar las necesidades y desafíos específicos de diferentes regiones del país.

A fin de llevar a cabo, el análisis de clusters fue necesario escalar los datos, para que una vez ya escalados poder llevar a cabo los diferentes estudios.

3.1. Cluster jerárquico aglomerativo AGNES

Este primer análisis de cluster fue generado con la función 'hclust' usando la medida de linkeo de enlace promediado.

A continuación, haciendo uso de la función cor, calculamos la correlación entre las distancias cophenetic del dendrograma y la matriz de distancias original.

El cálculo arrojó un valor de 0.8425647, un valor muy bueno, si consideramos que usualmente valores de 0.75 suelen considerarse buenos. Esto significa que las distancias en el dendrograma representan la verdadera similitud de las provincias de forma bastante realista y fiel.

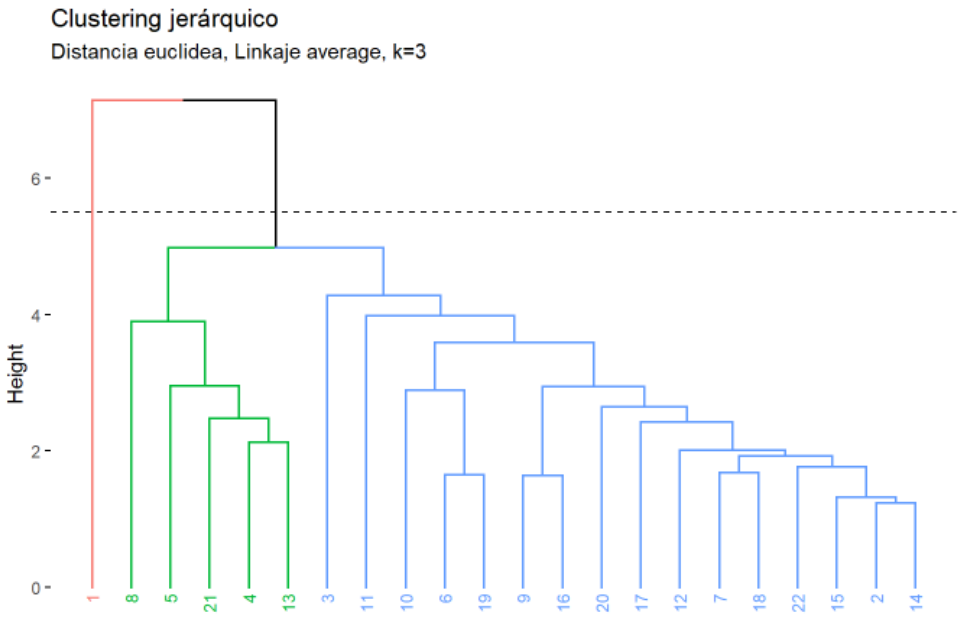

Figura 8

Llevamos a cabo la asignación de grupos con la función 'cutree' en el plano de las tres primeras componentes.

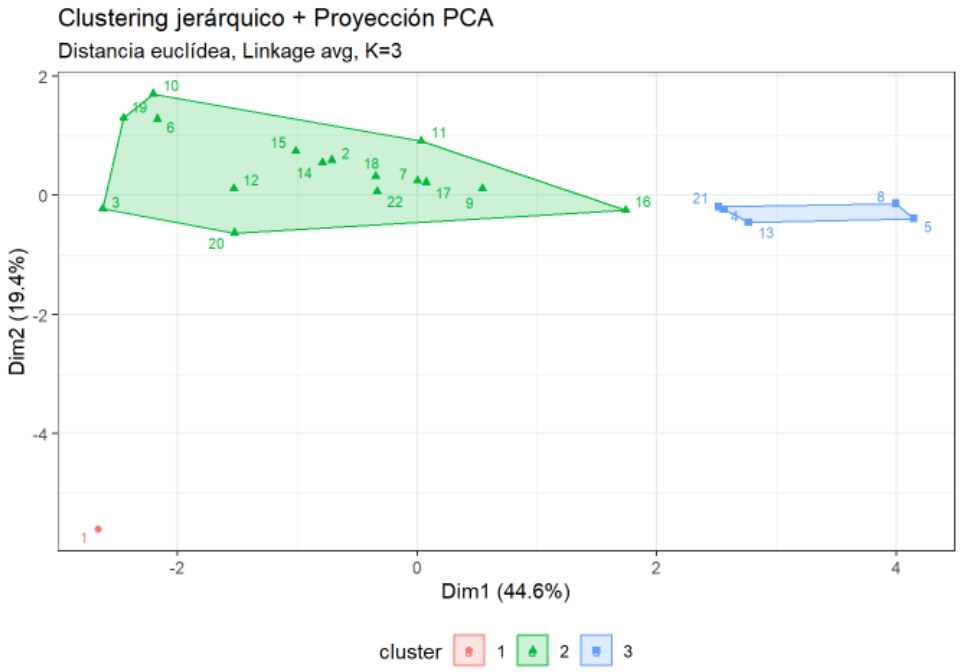

Figura 9

3.2. Cluster jerárquico divisivo DIANA

Ahora realizamos, de nuevo, un análisis de cluster jerárquico pero esta vez divisivo. Usamos la función 'diana' de la biblioteca 'cluster' y la distancia calculada entre las observaciones es la distancia euclídea.

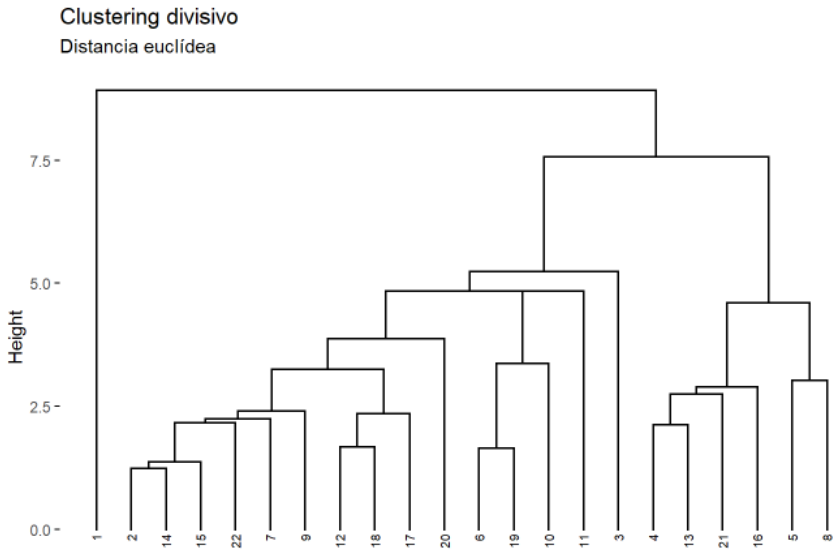

Figura 10

El mapa de calor (heatmap) muestra unos valores atípicos en las dimensiones de pbi y población para la provincia de Buenos Aires. El paquete usado fue 'viridis' y la función es 'heatmap'

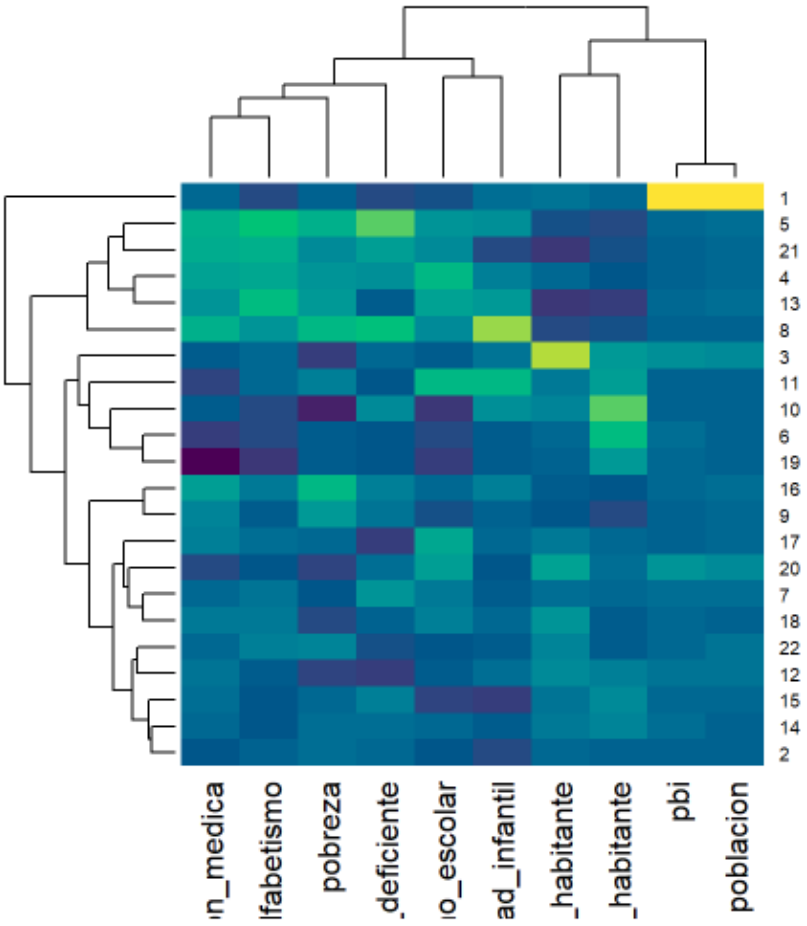

Figura 11: Heatmap

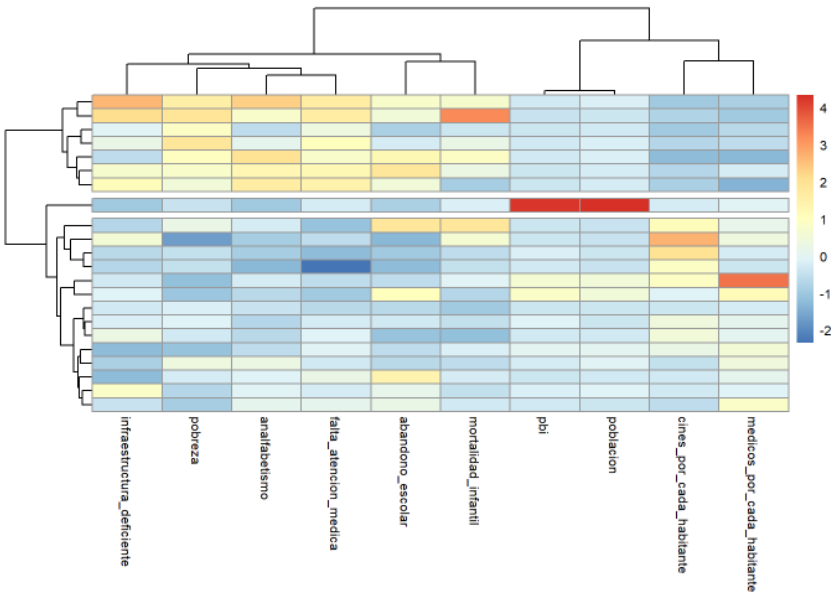

Figura 12: Pheatmap

3.3. Cluster no jerárquico k-means

Nuestra análisis ejecutando k-means sobre los datos arrojó información útil. Para ello usamos la función 'kmeans' con 3 centros.

Ahora buscamos establecer cuál es el número óptimo de clusters a crear.

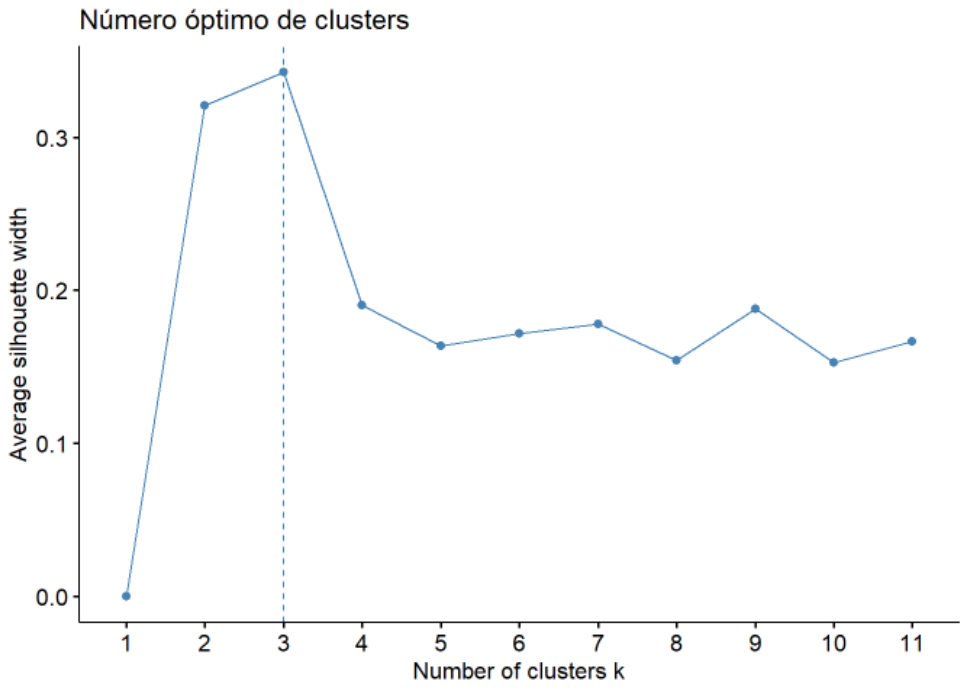

Figura 13: El número óptimo es de 3 clusters

Establecemos una semilla para reproducibilidad. Los resultados para k=3 fueron los siguientes:

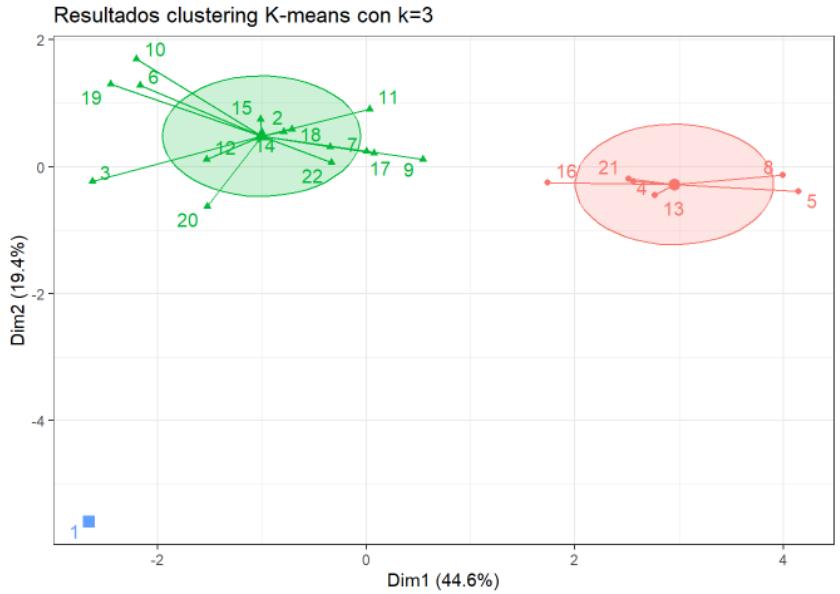

Figura 14

El análisis de silueta arroja un valor promedio de 0.34, un valor aceptable pero que da lugar a pensar

que los clusters no son muy consistentes ni demasiado compactos. La provicia de Buenos Aires muestra unos valores irrisorios.

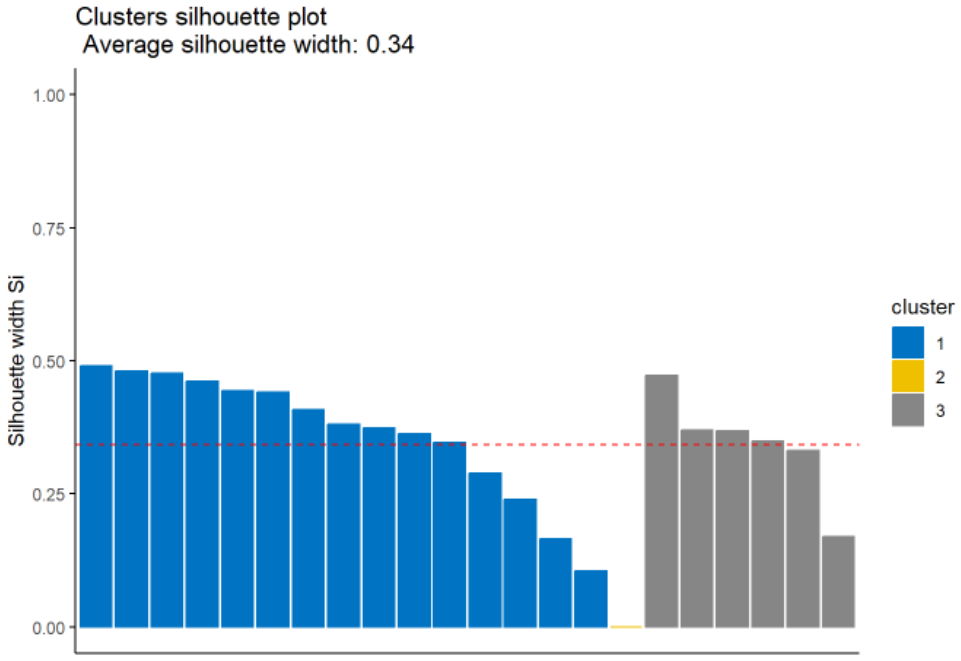

Figura 15

4. Conclusión

El análisis de componentes principales (ACP) y el clustering han revelado patrones distintivos en los datos provinciales de Argentina. Mediante el ACP, identificamos tres componentes principales que representan diferentes dimensiones de los indicadores socioeconómicos y de salud. Estos componentes capturan la variabilidad en los datos y nos permiten comprender mejor las relaciones entre las variables.

Por otro lado, el clustering agrupó a las provincias en tres categorías principales. En primer lugar, el 'Norte Grande' reúne a las provincias con los indicadores más bajos, caracterizadas por altas tasas de pobreza, analfabetismo y falta de acceso a servicios médicos. La 'Provincia de Buenos Aires' se destaca como un caso único debido a su tamaño y complejidad socioeconómica. Finalmente, el 'Resto del País' abarca un conjunto diverso de provincias que comparten similitudes en comparación con las otras dos categorías.

Estos resultados proporcionan una visión más profunda de la heterogeneidad provincial en Argentina. El ACP ha simplificado la complejidad de los datos al identificar las dimensiones clave, mientras que el clustering ha revelado patrones de agrupación significativos. Esta información es fundamental para la formulación de políticas y la asignación de recursos, ya que destaca las áreas que requieren una atención específica y permite una comprensión más precisa de la diversidad regional en el país.

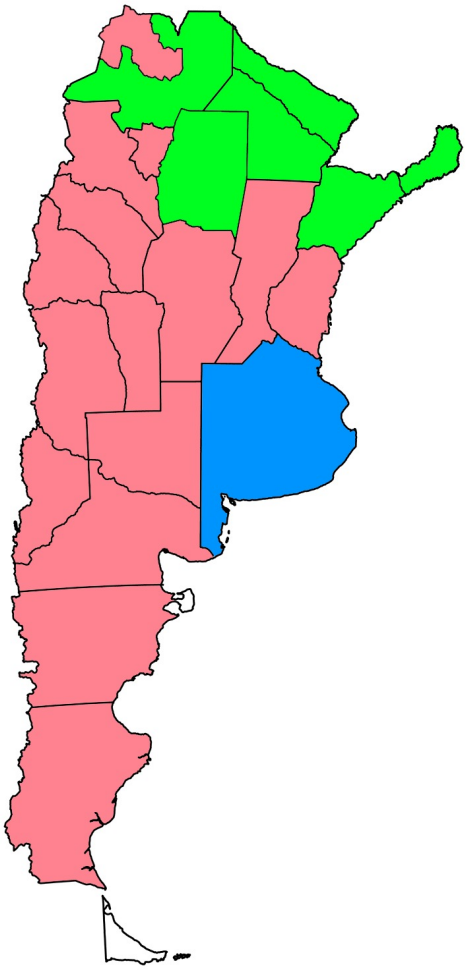

Figura 16: Mapa de Argentina con los 3 clusters de provincias.

5. Bibliografía

- Argentina provincial data
  https://www.kaggle.com/datasets/kingabzpro/argentina-provincial-data.