Big Data and Machine Learning: Trabajo Práctico 3.

Integrantes:

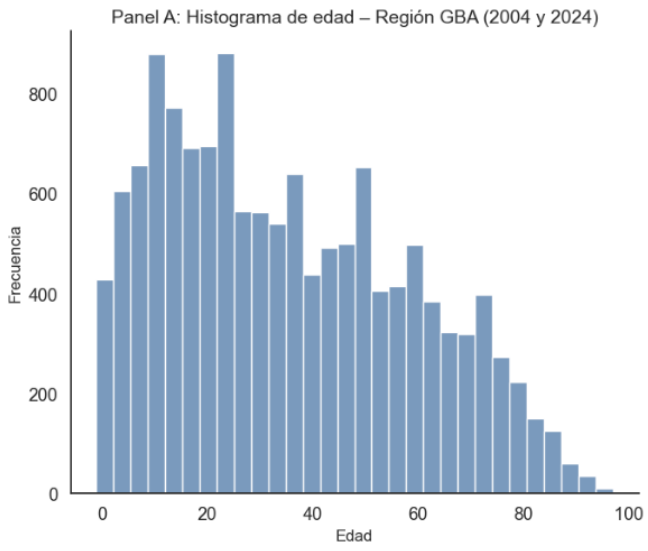
- Gabriela Yael Armoa (898589).
- Kevin Luque Benegas (915007).
- Carlos Ezequiel Martinez (911822).

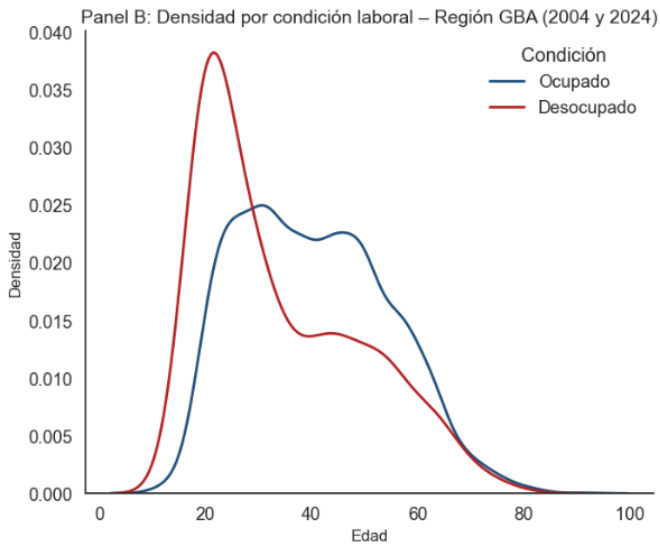
Parte I: Creación de variables, histogramas, kernels y resumen de la base de datos final.

1.

En el panel A, se observa un histograma que refleja la distribución de edades, con un pico pronunciado alrededor de los 20 años y una disminución gradual a medida que aumenta la edad. Esto sugiere que la mayoría de las observaciones se concentran en los grupos etarios más jóvenes.

Por otro lado, el panel B muestra la densidad de la variable edad, diferenciada por condición laboral. La línea azul representa a los ocupados, con un pico en torno a los 30-40 años, mientras que la línea roja, correspondiente a los desocupados, presenta un pico más marcado alrededor de los 20 años. Esto indica que los individuos desocupados tienden a ser más jóvenes en comparación con aquellos que se encuentran ocupados.





2.

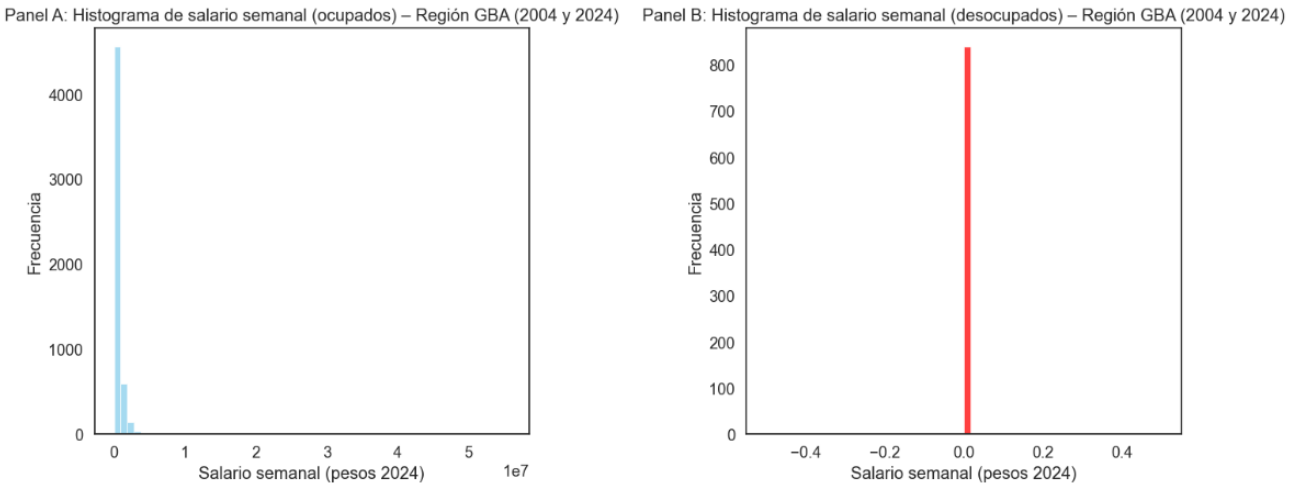
En la base de datos, el promedio de años de educación es de aproximadamente 10,98 años. Esto sugiere que, en su mayoría, las personas tienen un nivel educativo entre secundario incompleto o completo. El desvío estándar es de 4,69 años, lo que indica una gran heterogeneidad en los niveles educativos: algunos tienen más años de educación que el promedio, mientras que otros tienen menos.

El valor mínimo es 0, lo que refleja que algunas personas no han recibido educación formal, posiblemente debido a factores como edad o condiciones socioeconómicas. Por otro lado, la mediana es de 13 años, que es mayor que el promedio, quiere decir que la distribucion tiene una simetria por la izquierda. Lo que significa que la mitad de la población tiene 13 o más años de educación, lo que equivale a secundario completo o algo más. Finalmente, el valor máximo de 23 años indica que hay personas con un alto nivel educativo, como aquellos con posgrados o múltiples carreras universitarias.

	Educación (años)
Promedio	10.979245
Desvío estándar	4.691244
Mínimo	0.000000
Mediana (P50)	13.000000
Máximo	22.000000

3.

Los paneles muestran la distribución de salarios semanales para ocupados y desocupados. Vemos que los ocupados tienen mayores salarios que los desocupados, por evidentes razones de percibir un ingreso laboral que es fijo y por varias horas. Esto refleja la precaria situación económica de los desocupados.



4.

En cuanto a las estadísticas de las horas trabajadas, se observa que el promedio es bajo (16.2 horas), lo que podría indicar que una gran parte de los individuos en la muestra trabaja pocas horas o no trabajan en absoluto. Esto se ve reflejado en la mediana (0 horas), lo que sugiere que la mitad de los casos no reportan actividad laboral, sea porque no trabajan o no quieren decir cuánto.

La desviación estándar es elevada (50.8), lo que denota una alta dispersión en las horas trabajadas, con algunos casos extremos de trabajadores que han registrado hasta 1047 horas, lo que podría ser un error en los datos o situaciones excepcionales como trabajos extraordinarios.

Este patrón podría indicar una segmentación en el mercado laboral, con una gran proporción de trabajadores a tiempo parcial o inactivos y algunos pocos con una carga laboral mucho mayor. Esto podría ser relevante al analizar la informalidad laboral o las condiciones de empleo en el contexto estudiado.

```
Promedio: 16.198734729493893
Desviación estándar: 50.78792586974127
Mínimo: 0.0
Percentil 50 (Mediana): 0.0
Máximo: 1047.0
```

5.

En el trabajo práctico realizado, se observa que la muestra total consta de 13.752 observaciones, distribuidas en 7.647 correspondientes al año 2004 y 6.105 al año 2024. Cabe destacar que no se registraron valores faltantes (NAs) en la variable "Estado".

Respecto a la condición laboral, se contabilizaron 5.357 ocupados y 839 desocupados en el total de la muestra. Al analizar la evolución entre los dos años, se aprecia una disminución tanto en la cantidad de ocupados como de desocupados.

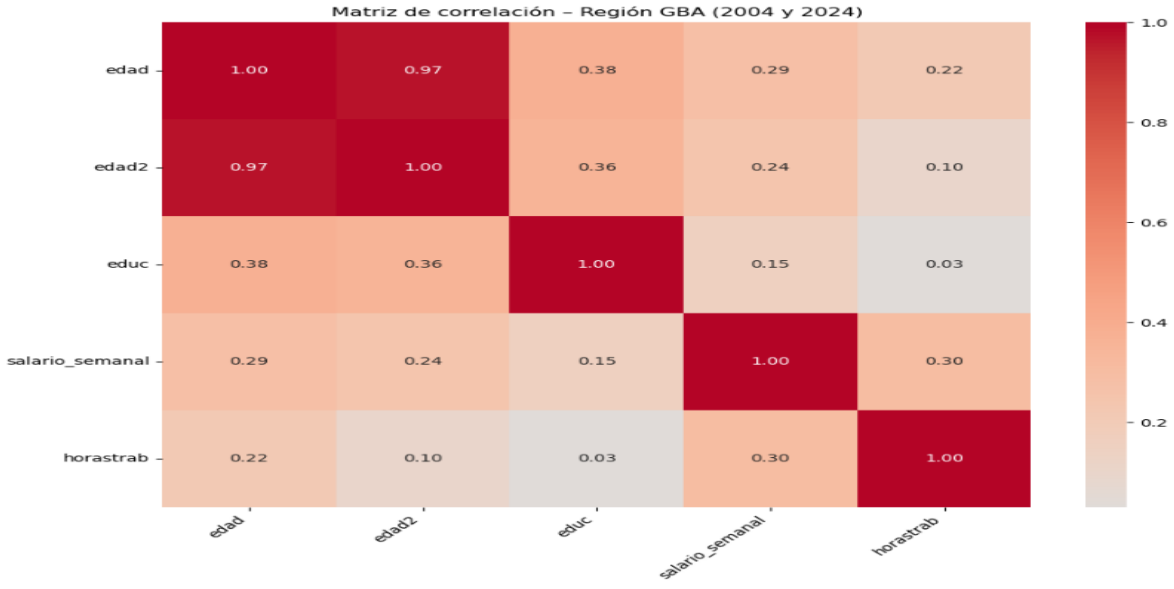
Por último, se indica que se han limpiado y homogeneizado 24 variables para el año 2004 y 18 variables para el 2024, finalizando con 16 variables comunes a ambos períodos.

	2004	2024	Total
Cantidad observaciones	7647	6105	13752
Cantidad de observaciones con Nas en la variable "Estado"	0	0	0
Cantidad de Ocupados	3079	2278	5357
Cantidad de Desocupados	528	311	839
Cantidad de variables limpias y homogeneizadas	24	18	16

Parte II: Métodos No Supervisados.

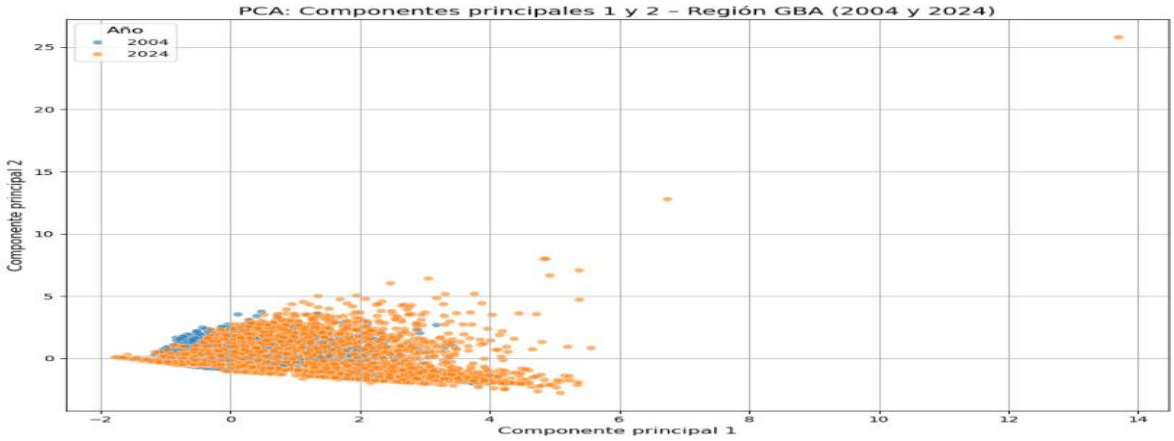
1.

Podemos ver una matriz de correlación entre cinco variables seleccionadas: edad, edad², años de educación, salario semanal y horas trabajadas. Podemos observar que, edad y edad² están casi perfectamente correlacionadas porque una es el cuadrado de la otra. También se puede ver que, en las variables de salario semanal y horas trabajadas, existe una correlación positiva, a mayores horas trabajadas, más gana una persona. Después tenemos las variables de educación y salario, las cuales aparentemente no se relacionan tanto, es decir, estudiar más no quiere decir que vayas a ganar mucha más plata según los datos.



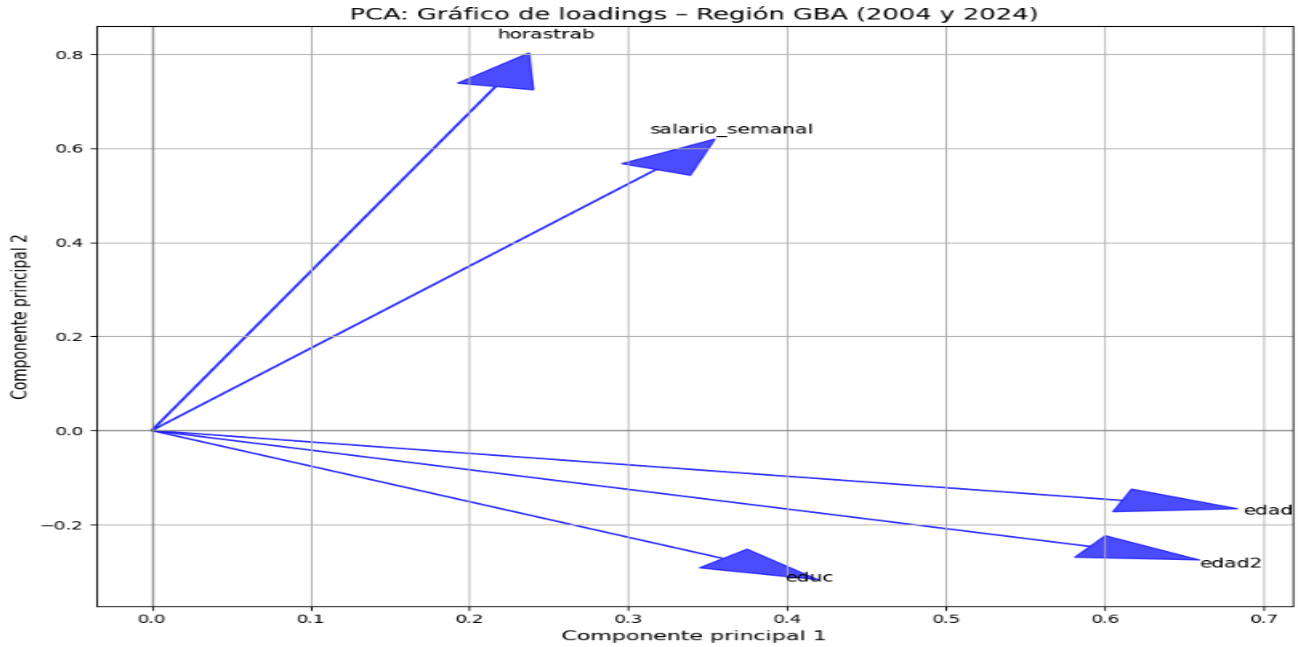
2.

En el gráfico muestra los resultados del PCA, reduciendo las cinco variables originales a dos componentes que explican la mayor parte de la varianza. Se puede observar una cierta separación entre los dos grupos, lo que sugiere que hubo cambios estructurales entre ambos años. La dispersión en el espacio PC1-PC2 permite visualizar cómo se agrupan los datos.



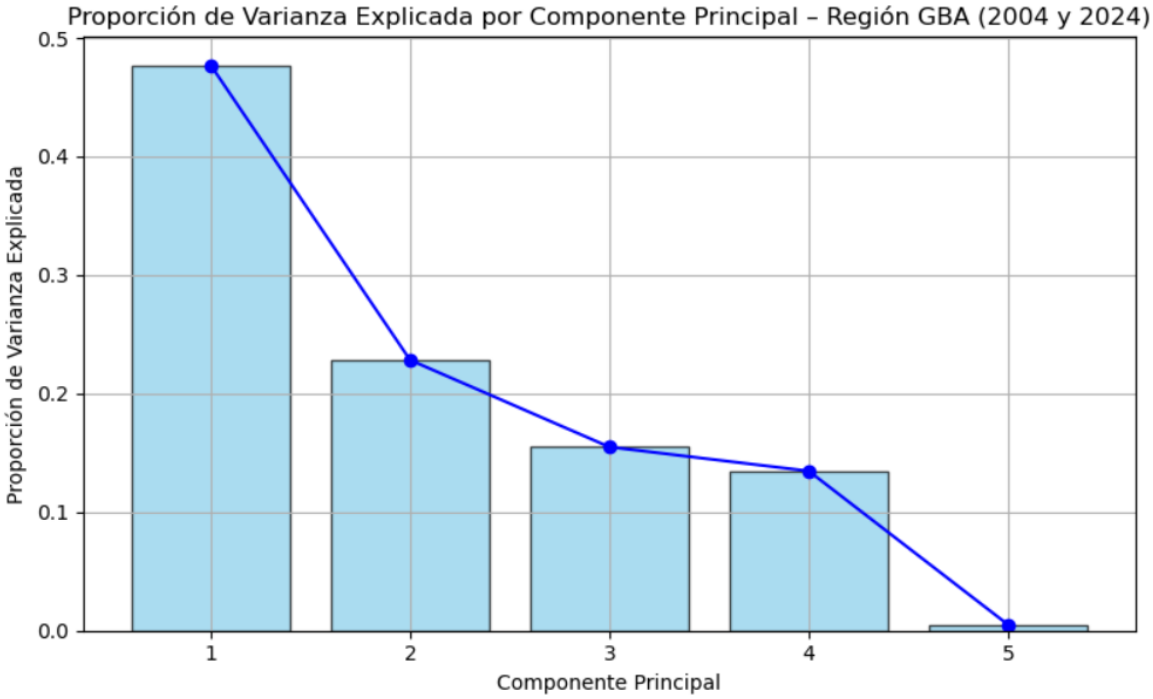
3.

En el gráfico se visualizan los loadings del PCA. Las flechas indican la dirección e intensidad de la contribución de cada variable. Se destaca que edad y edad² tienen una mayor influencia sobre PC1, mientras que horas trabajadas y salario semanal tienen mayor peso en el PC2. El gráfico permite interpretar cómo cada variable ayuda a estructurar el espacio reducido de análisis.



4.

En el gráfico podrán observar la proporción de varianza explicada de cada componente y su representación en valores del 0 al 0.5, en el que se explica el tamaño que tiene cada uno según la base de datos aplicando el PCA.

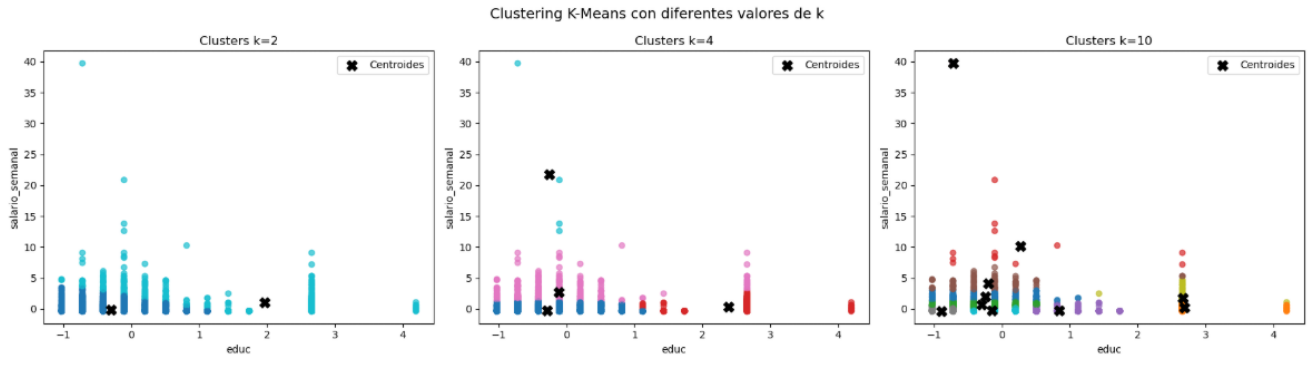


5.

A-)

En los gráficos podemos ver los cambios que provoca el valor “k” en los valores de salario semanal x educ y sus respectivos centroides, con estos datos podemos interpretar que a mayor k=x más dispersión y más centroides aparecerán en el gráfico, y en cuanto a los gráficos podemos entender que:

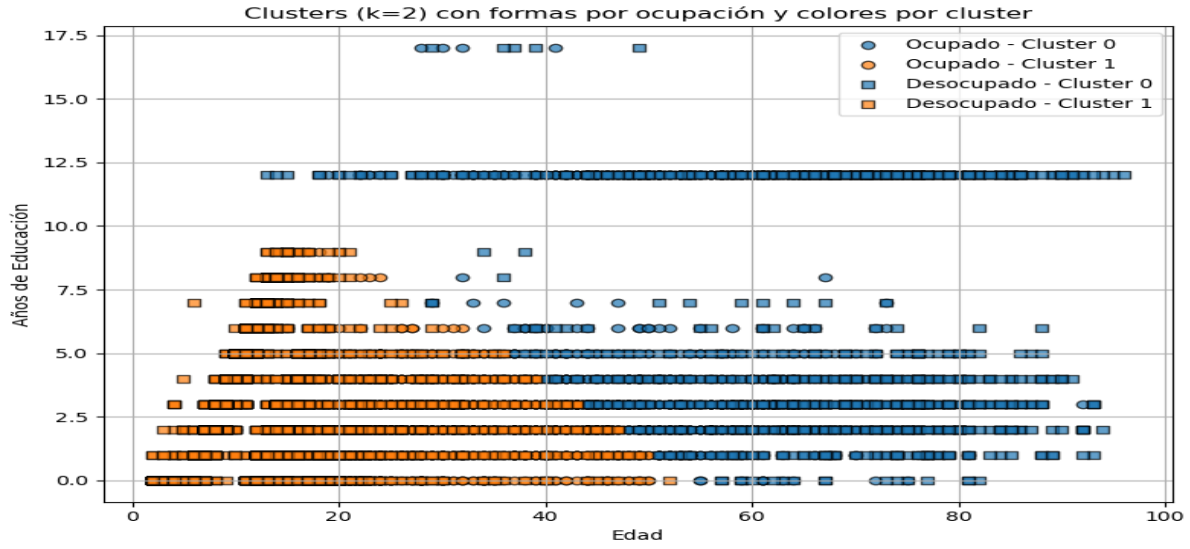
- Con **k=2**: El algoritmo separa la población en dos grandes grupos (por ejemplo, baja vs. alta educación/salario).
- Con **k=4**: Se distinguen más subgrupos, posiblemente reflejando niveles educativos y salariales más matizados.
- Con **k=10**: Aparecen muchos grupos, lo que puede reflejar sobreajuste si no hay tanta diferencia real entre los subgrupos.



B-) No se puede ya que: (Primer código que se ve en la parte B)

- Si los ocupados y desocupados caen en diferentes clusters, el modelo estaría capturando **alguna diferencia real** entre ambos grupos.
- Si están mezclados, entonces el clustering **no logra distinguir bien entre ocupados y desocupados** solo con edad y educación.

  Es por esto que vamos a usar la aproximación asi podemos obtener algunos datos correctos en su mayoría, pero no son del todo confiables por lo que recomendamos no tener como algo 100% representativo al grafico por aproximación y lo podremos ver en la figura 12

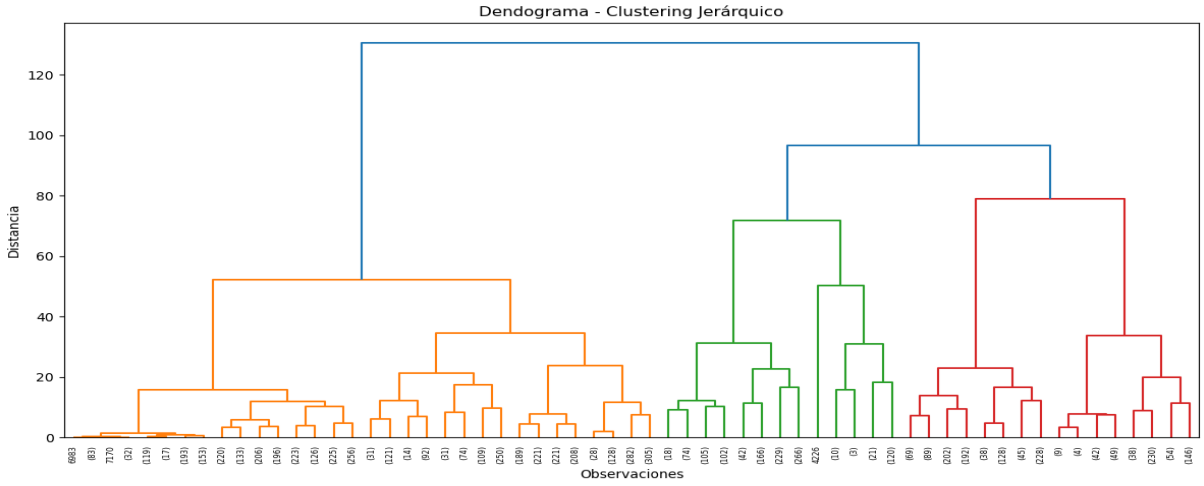


6.

Un **dendograma** es un diagrama en forma de árbol que muestra cómo se agrupan progresivamente las observaciones en un análisis de clustering jerárquico.

Cada rama representa una unión entre clusters, y la **altura** a la que se unen indica la **distancia (diferencia)** entre ellos. Cuanto más alta sea la unión, **menos similares** son los grupos.

En el último gráfico vemos, por último, los datos finales que hemos obtenido al hacer un dendograma con todos los datos y observaciones según su distancia.


Dendograma - Clustering Jerárquico
Distancia
0
20
40
60
80
100
120
Observaciones