BIG DATA & MACHINE LEARNING

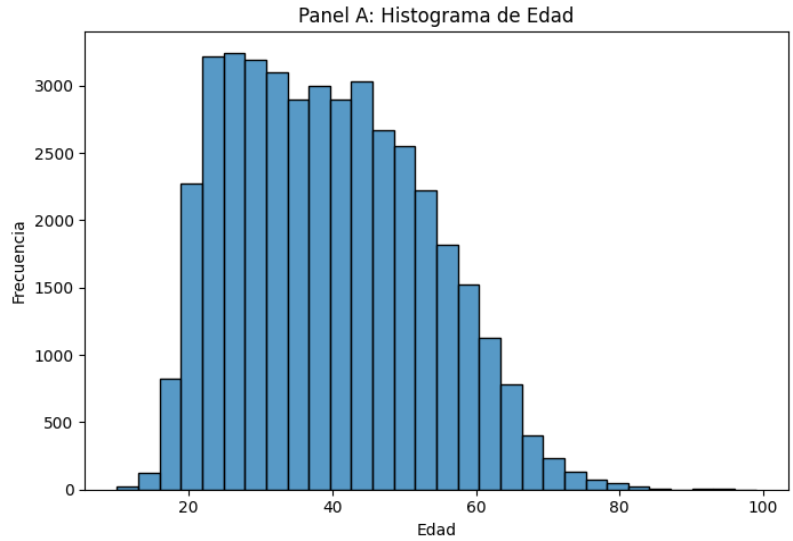
TRABAJO PRÁCTICO N° 3

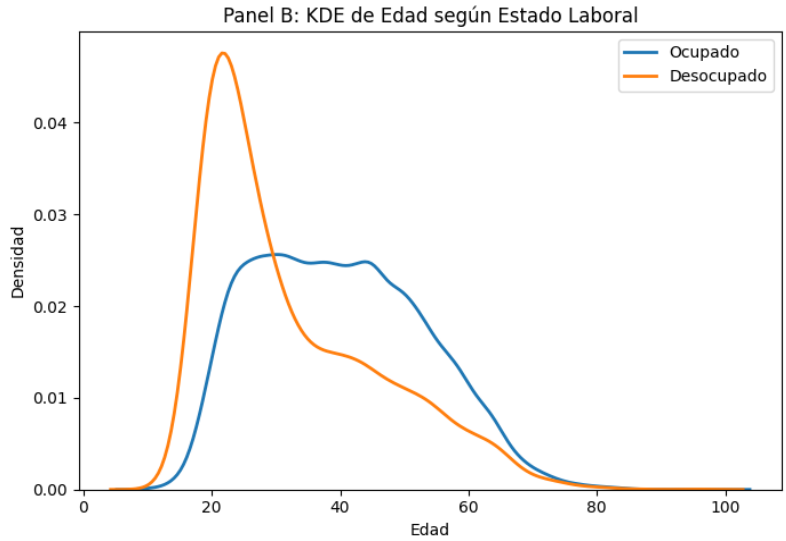
HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

FACUNDO HERRERO

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

1-1)





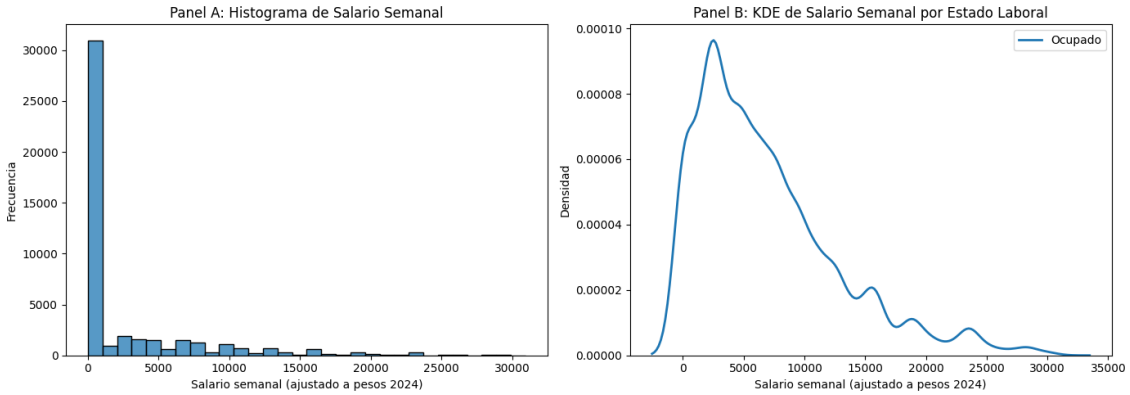
En el Panel A se deja ver que la mayoría de los encuestados tiene entre 20 y 50 años, con una caída progresiva hacia edades mayores. El Panel B muestra que los desocupados tienden a concentrarse en edades más jóvenes, especialmente entre los 20 y 25 años, mientras que entre los ocupados la distribución es más pareja y se extiende a edades mayores. Esto podría reflejar que el desempleo afecta con más fuerza a los jóvenes, también que los adultos tienen una mayor inserción en el mercado laboral.

1-2)

ESTADÍSTICA DESCRIPTIVA	Valor
Observaciones	958
Promedio	9,83
Desvío estándar	4,52
Mínimo	0
Mediana (p50)	11
Máximo	18

Se construyó la variable ‘educ’ como cantidad de años de educación formal a partir de las variables CH12, CH13 y CH14. La mediana es de 11 años, con un promedio de 9.8 y un máximo de 18, lo que refleja una mayoría con secundario incompleto o completo y menor presencia de niveles terciarios o universitarios.

1-3)



En el Panel A se observa que una gran proporción de personas reporta un salario semanal igual a cero, reflejando la situación de los desocupados o personas sin ingresos laborales. Para quienes sí perciben ingresos, la distribución es muy asimétrica, con una mayoría concentrada en niveles bajos de salario y una cola larga hacia ingresos más altos. En el Panel B, la curva de densidad confirma esta distribución desigual, con una mayor densidad en tramos bajos y una rápida caída a medida que aumenta el salario.

1-4)

Estadística descriptiva

	Valor
Observaciones	404
Promedio	38.54
Desvío estándar	16.15
Mínimo	1
Mediana (p50)	40.0
Máximo	84

Se creó la variable ‘horastrab’ sumando las horas semanales trabajadas en la ocupación principal y secundaria. Se filtraron los casos para incluir solo personas ocupadas que declararon entre 1 y 84 horas semanales, eliminando registros nulos. Luego del filtrado, se obtuvieron 404 observaciones válidas. La cantidad promedio de horas trabajadas fue de 38,5 horas semanales, con una mediana de 40 horas, lo que indica una fuerte concentración en jornadas laborales normales. También se observó una mínima dispersión hacia empleos de menor carga horaria, con valores que oscilan entre 1 y 84 horas.

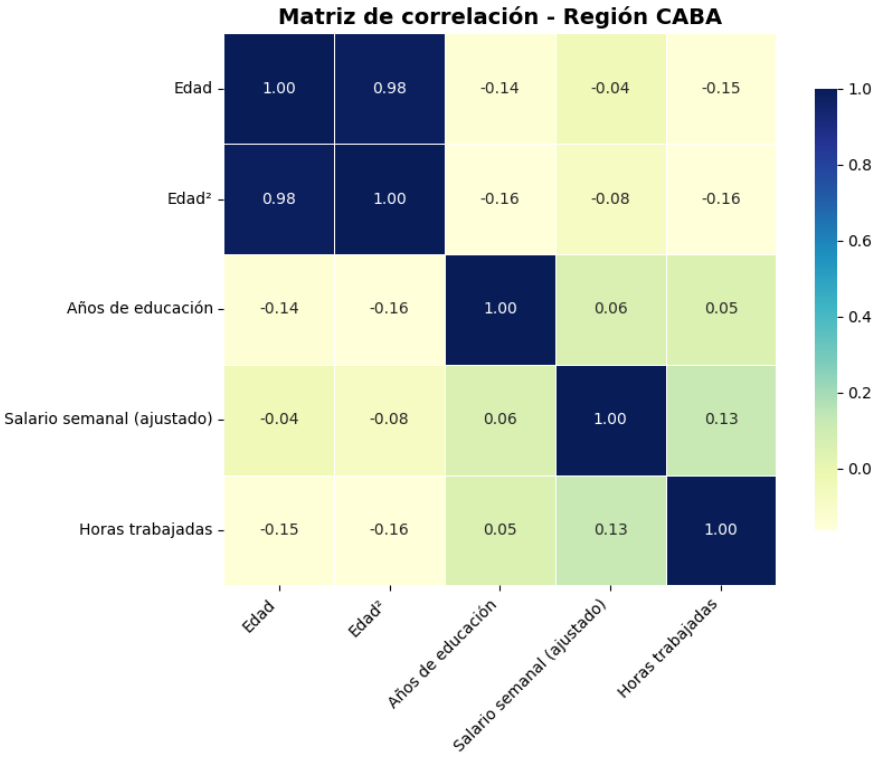
1-5)

La base final para la región Ciudad Autónoma de Buenos Aires (CABA) contiene 3.258 observaciones, provenientes de las bases individuales de 2004 y 2024. No se registran valores faltantes en la variable “estado”, y se identificaron 1.595 ocupados y 170 desocupados en total. Las variables utilizadas fueron correctamente unificadas, resultando en tres variables limpias y homogéneas entre ambos años.

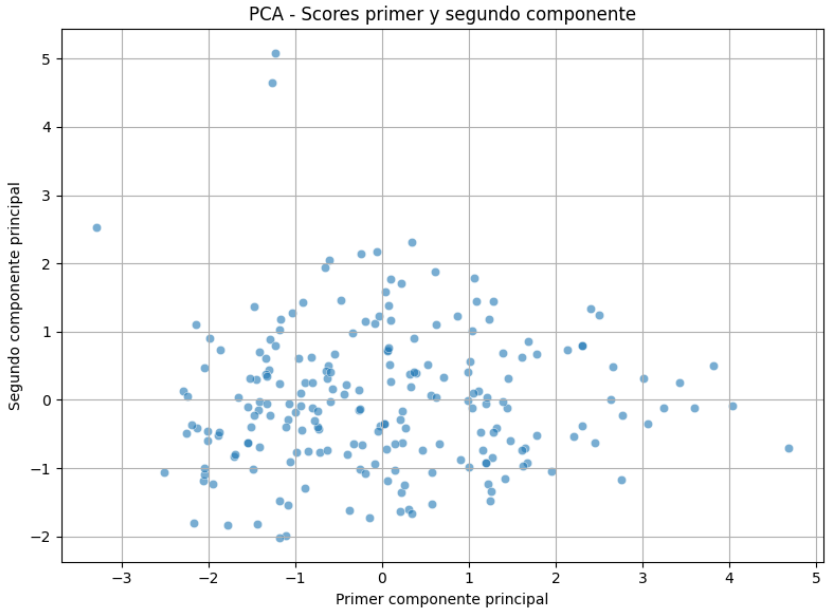
	2004	2024	Total
Cantidad de observaciones	1836	1422	3258
Cantidad de NaNs en 'estado'	0	0	0
Cantidad de ocupados	838	757	1595
Cantidad de desocupados	119	51	170
Variables limpias y homogeneizadas	3	3	3

Parte II: Métodos No Supervisados

1)



2)



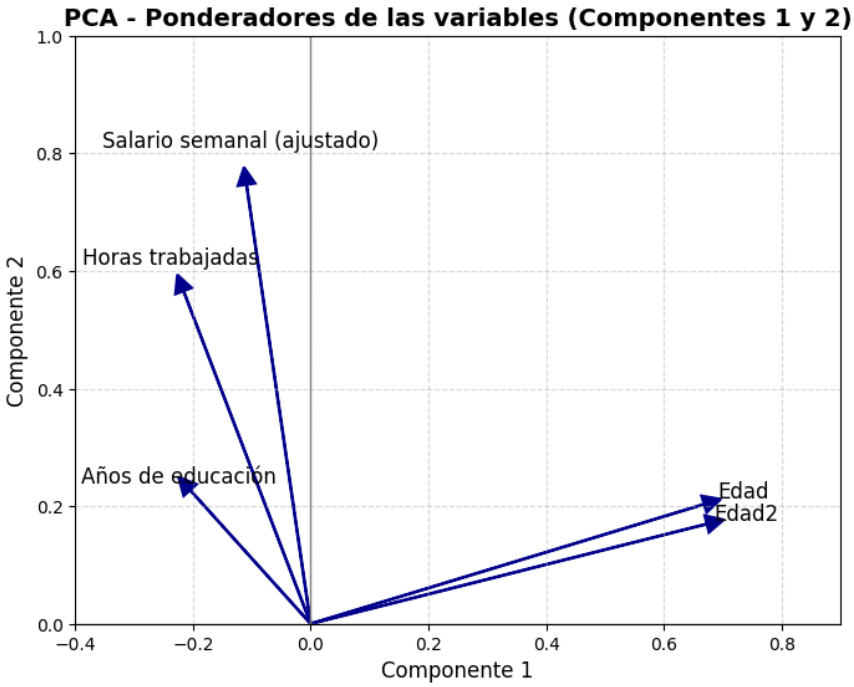
Varianza explicada por cada componente:

PC1: 41.78%

PC2: 21.68%

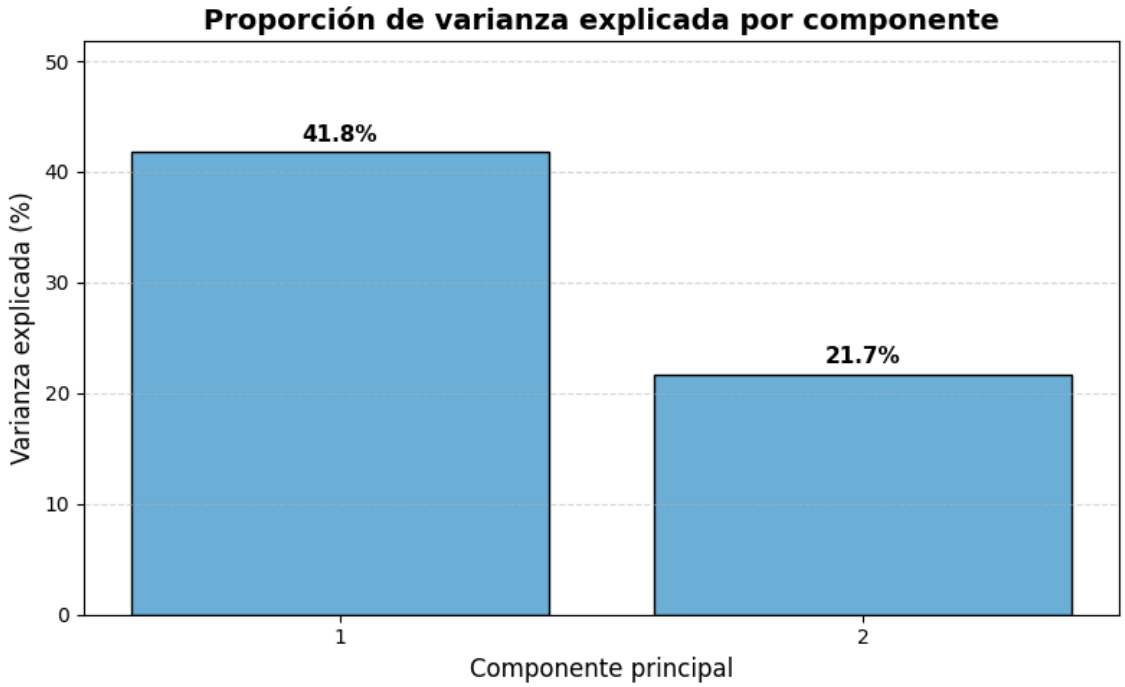
Los dos primeros componentes principales del análisis explican en conjunto el 63,46 % de la variabilidad del conjunto de datos, lo cual representa un nivel razonable de síntesis. El primero de ellos concentra más del 40 % de la información, lo que permite reducir la dimensionalidad sin perder demasiado contenido relevante. Al observar el gráfico de dispersión, se nota que los puntos están bastante concentrados alrededor del centro y no se distinguen agrupamientos definidos, lo que indica que los componentes capturan bien la estructura general, pero no permiten identificar patrones claramente diferenciados entre individuos.

3)



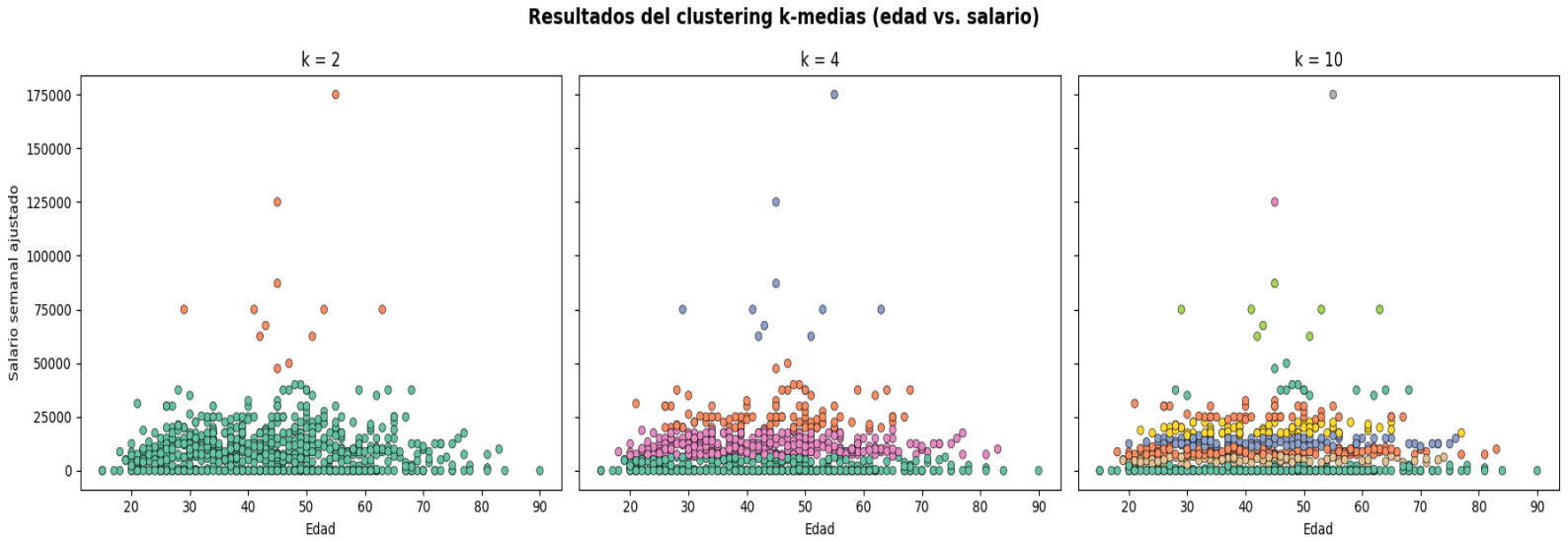
En el gráfico se observan los ponderadores (loadings) de cada variable sobre los dos primeros componentes principales del PCA. La variable Edad (y su cuadrado) tiene una alta carga positiva sobre el primer componente, indicando que este eje está fuertemente asociado al perfil etario. En cambio, salario semanal (ajustado) y horas trabajadas presentan cargas altas en el segundo componente, lo que sugiere que ese eje refleja más bien condiciones del mercado laboral. Años de educación se encuentra en una posición intermedia, con cargas moderadas en ambos ejes. Este análisis permite interpretar que los dos componentes resumen dimensiones distintas: el primero, vinculada a la edad, y el segundo, a la inserción laboral.

4)



Cómo ya comenté anteriormente, el gráfico muestra la proporción de varianza explicada por los dos primeros componentes principales del análisis PCA. El primer componente explica un 41.8% de la variabilidad de los datos, mientras que el segundo explica un 21.7% adicional. Esto significa que, en conjunto, los dos componentes retienen más del 63% de la información original. Esta distribución también indica que el primer componente captura mucho más de la estructura interna del conjunto de datos que el segundo.

5)



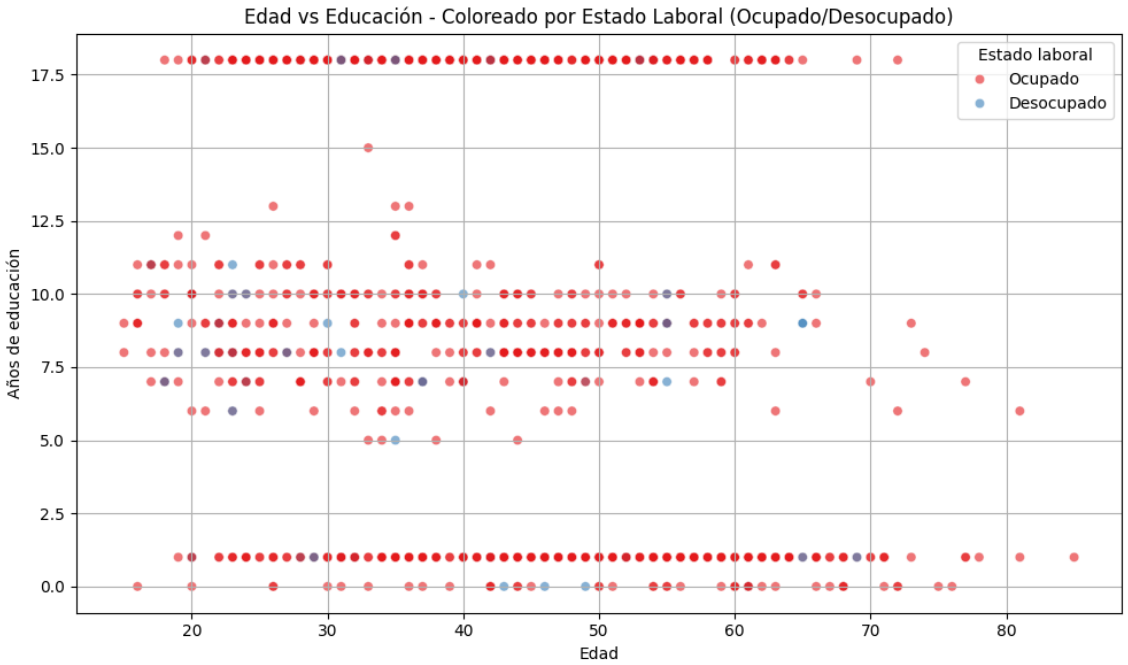
La visualización del clustering k-medias con k = 2, 4 y 10 permite observar cómo se agrupan los individuos según su edad y salario semanal ajustado.

Con k = 2, el algoritmo distingue un grupo mayoritario con salarios bajos y otro minoritario con ingresos elevados, sin hacer distinciones claras por edad. Al aumentar a k = 4, aparecen agrupamientos más definidos, combinando características de edad e ingresos, como adultos

mayores con bajos ingresos y personas de mediana edad con ingresos medios o altos. Con k = 10, los grupos se vuelven más específicos y numerosos, lo que puede dificultar su interpretación general y sugiere una segmentación demasiado detallada.

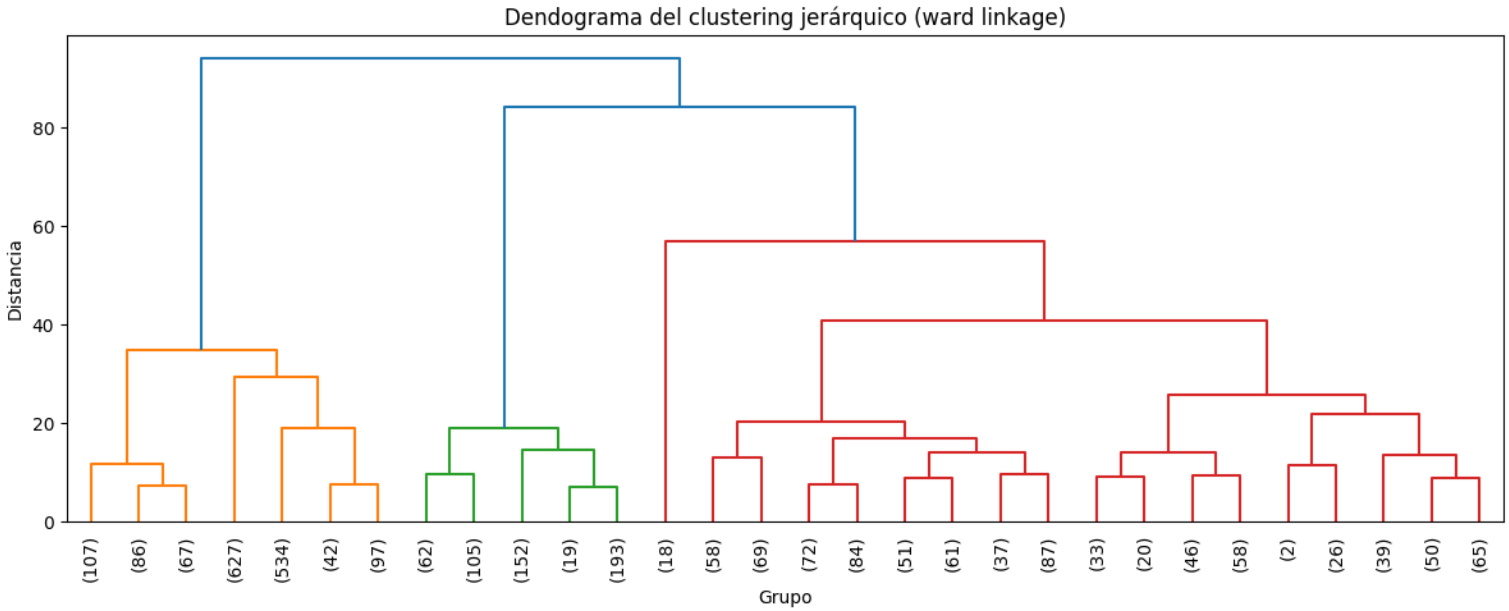
En resumen, la variable salario parece tener mayor peso en la formación de los clusters, y valores intermedios de k, como 4, ofrecen un buen equilibrio entre detalle e interpretabilidad.

5b)



El algoritmo no logró separar correctamente a las personas ocupadas de las desocupadas. En el gráfico se observa una superposición entre ambos grupos en casi todos los niveles de educación y edades. Si bien hay ciertas tendencias, como mayor concentración de ocupados con más educación, no existe una división clara que el modelo de k-means pueda capturar solo con estas variables.

6)



Un dendograma es un gráfico en forma de árbol que muestra cómo se agrupan progresivamente los datos en un análisis de clustering jerárquico. Cada unión entre ramas representa la combinación de grupos, y la altura indica cuán diferentes eran entre sí al momento de unirse. El análisis jerárquico de este dendograma permite agrupar a las personas según características como edad, educación, salario y horas trabajadas, sin necesidad de indicar cuántos grupos usar desde el inicio. El gráfico del endograma muestra cómo se fueron uniendo los grupos paso a paso. Se puede ver que, al menos hasta formar 3 o 4 grupos, las uniones se hacen con diferencias moderadas, lo que sugiere que esos grupos tienen sentido. A diferencia de otros métodos, este enfoque permite ver cómo se relacionan todos los individuos y cómo se fueron formando los grupos. La altura a la que se unen los grupos indica qué tan distintos eran entre sí. En este caso, elegir 3 o 4 grupos parece una buena opción para analizar los datos sin perder detalle.