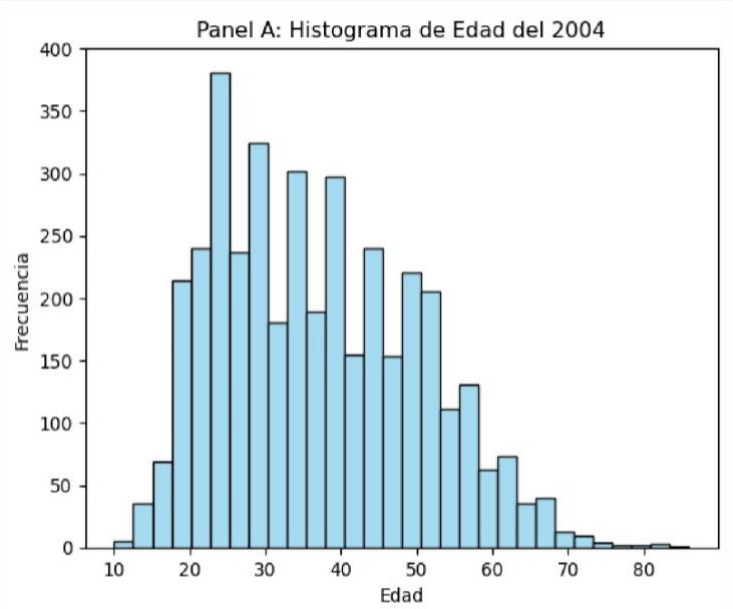
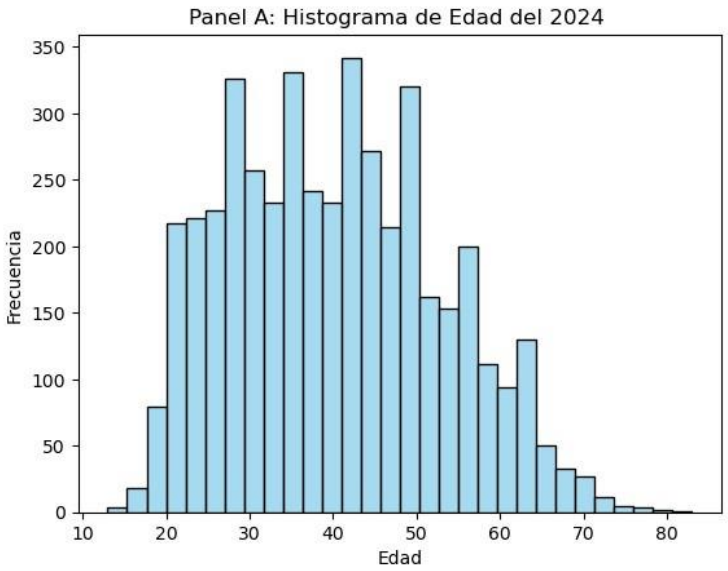
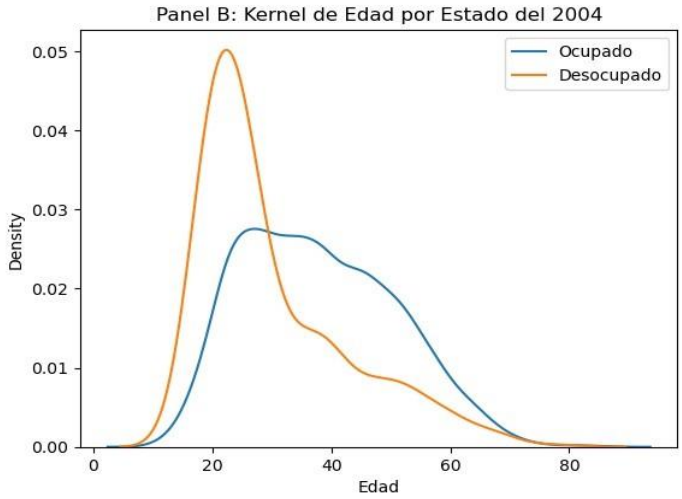
Big Data y Machine Learning (UBA) - 2025

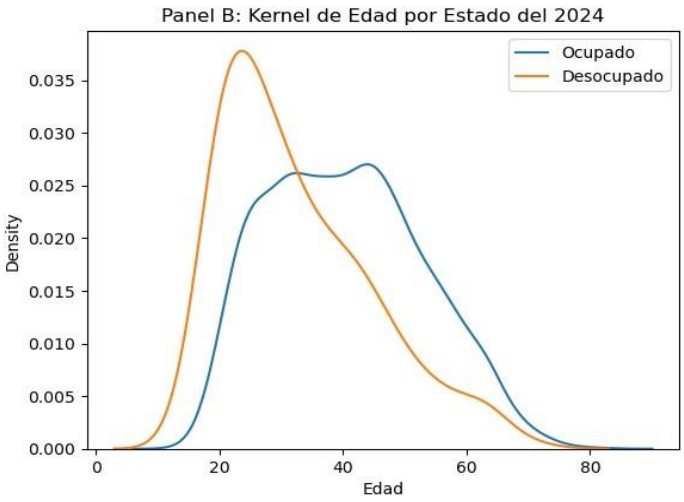
Trabajo Práctico 3: HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

Grupo: 10
Integrantes: Bautista Benetti, Luca D'adderio y Tongkun Weng







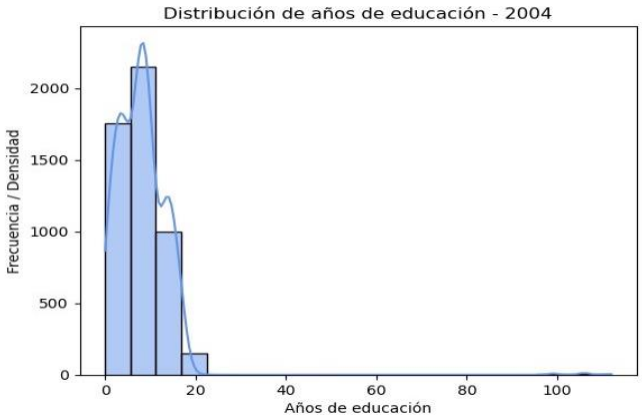


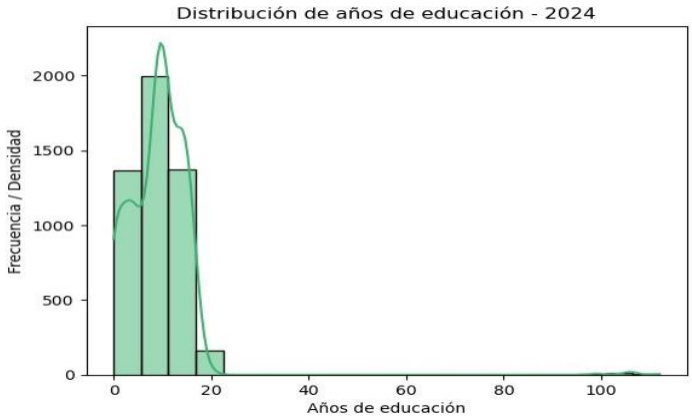
Hay evidencia de un envejecimiento de la población o al menos un desplazamiento de la edad modal hacia grupos mayores en 2024.
La desocupación sigue afectando desproporcionadamente a los jóvenes en ambos años.
La distribución de los ocupados es más homogénea en 2024, lo que podría reflejar mejoras en la empleabilidad de adultos mayores o cambios en el mercado laboral.

Estadísticas descriptivas del 2004:

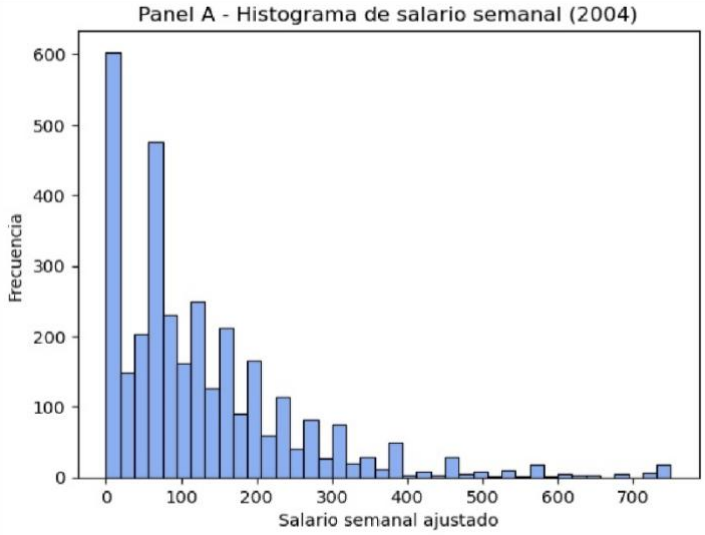
count	5064.000000
mean	8.118483
std	7.031895

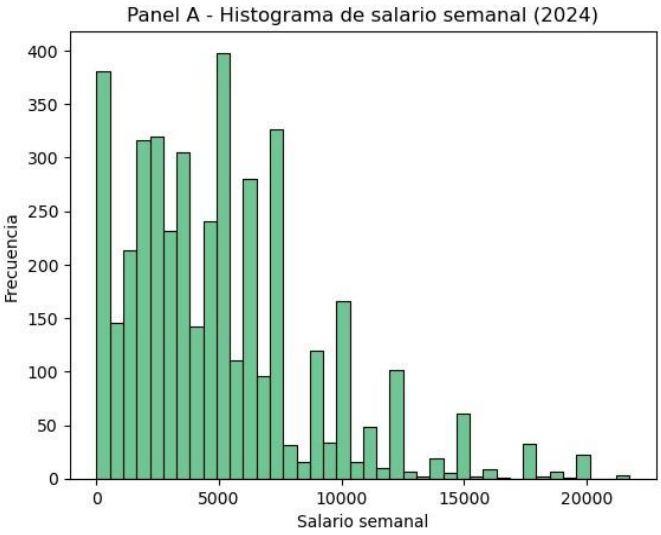
```
min          0.000000
50%          8.000000
max        112.000000
Name: educ, dtype: float64
Estadísticas descriptivas del 2024:
count     4920.000000
mean         9.297561
std          8.122692
min          0.000000
50%          9.000000
max        112.000000
Name: educ, dtype: float64
```

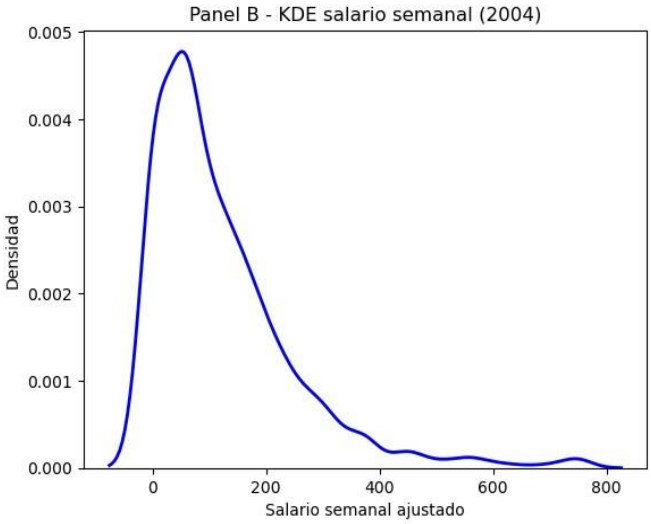


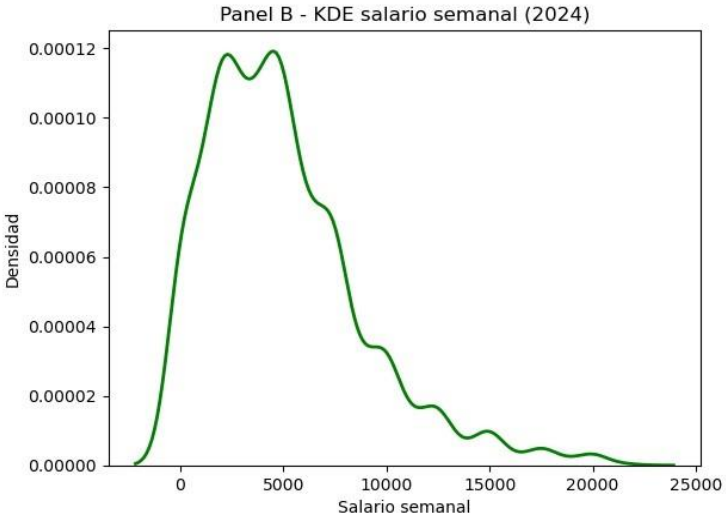


Mejora educativa: La población en 2024 parece tener más años de educación promedio que en 2004. Esto es consistente con una mejora del acceso a la educación o mayores niveles de finalización.
Persistencia de baja escolaridad: A pesar del avance, muchos individuos aún tienen niveles bajos de educación, especialmente evidenciado por la alta frecuencia en los rangos de 0 a 8 años.
Distribución similar: Aunque hay una leve mejora en la media, las formas de ambas distribuciones son bastante similares, lo que indica que no ha habido un cambio radical en la estructura educativa de la población.







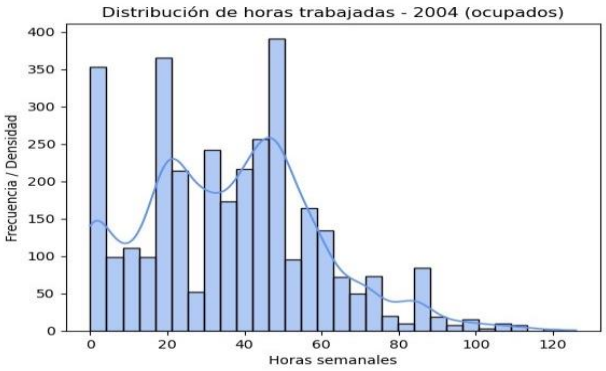


Aumento generalizado del salario nominal: En 2024 los ingresos semanales son mucho más altos en términos absolutos que en 2004, lo cual puede reflejar inflación, crecimiento económico o ambos.
Persistencia de desigualdad: A pesar del aumento en los salarios, la distribución sigue siendo asimétrica, lo que indica que una porción importante de la población sigue ganando por debajo de la media.
Mayor dispersión: Hay más individuos con ingresos altos en 2024, lo que podría sugerir un ensanchamiento de la brecha salarial.

Estadísticas de horas trabajadas - Año 2004 (ocupados)

count	3337.000000
mean	36.326940
std	23.008081
min	0.000000
25%	20.000000
50%	36.000000
75%	50.000000
max	126.000000

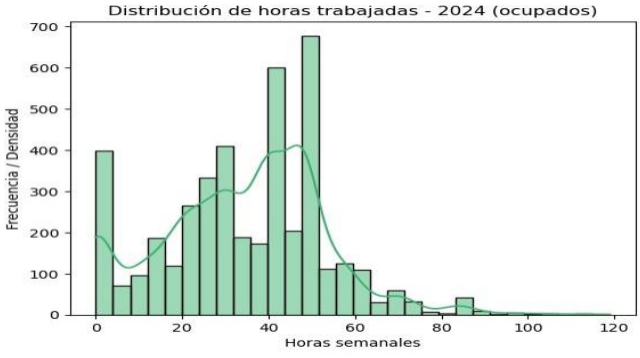
Name: horastrab, dtype: float64



Estadísticas de horas trabajadas - Año 2024 (ocupados)

count	4263.000000
mean	33.973258
std	18.947363
min	0.000000
25%	20.000000
50%	36.000000
75%	48.000000
max	119.000000

Name: horastrab, dtype: float64



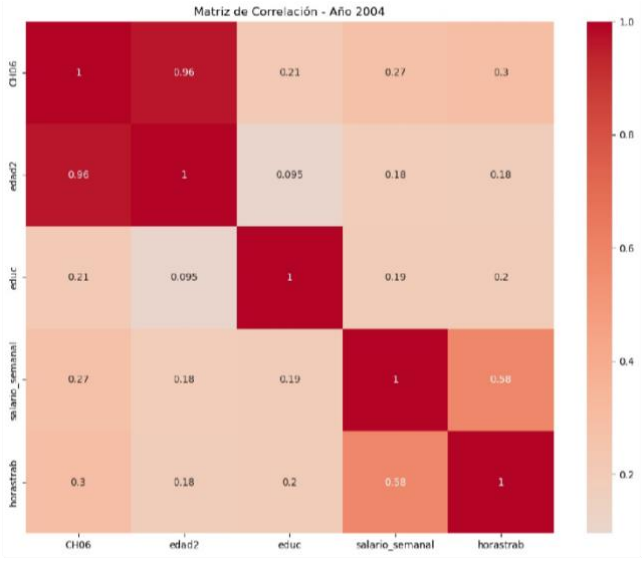
Persistencia de la jornada estándar de 48 horas: En ambos años, hay un pico en torno a las 48 horas. Esto sugiere que hay estabilidad en el régimen de

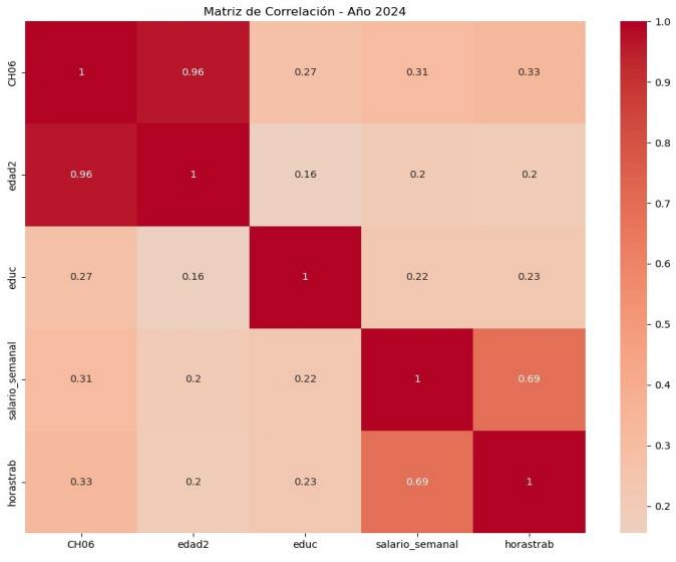
trabajo formal.
Leve disminución en promedio de horas trabajadas: Aunque la mediana sigue en 36, la media bajó casi 2.5 horas, lo que puede indicar cambios en la estructura laboral (más trabajo parcial, menos horas extras, o más heterogeneidad).
Mayor concentración en 2024: Hay menos dispersión en la distribución de 2024, lo que indica una cierta homogeneización de las jornadas laborales.
Presencia constante de trabajadores subocupados: Hay una franja significativa que trabaja pocas horas por semana en ambos años, lo que podría reflejar subocupación o informalidad persistente.

Tabla 1. Resumen de la base final para la región 40

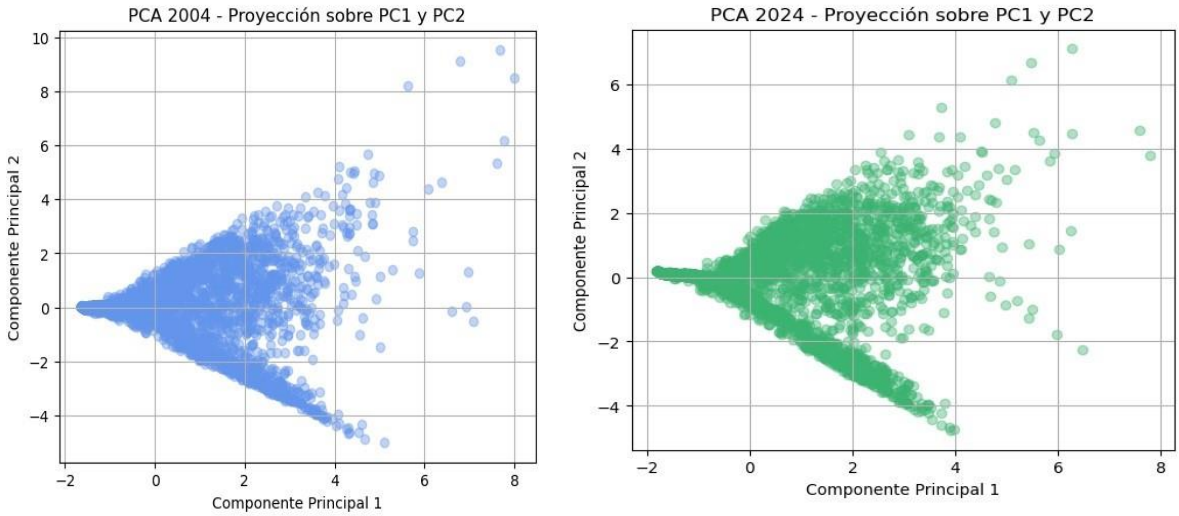
	2004	2024	Total
Cantidad observaciones	9393	9699	19092
Cantidad de observaciones con Nas en las variable “Estado”	8	0	8
Cantidad de Ocupados	3337	4263	7600
Cantidad de Desocupados	597	256	853
Cantidad de variables limpias y homogeneizadas	5058	4920	9978

El tamaño total de la base de datos unificada para la región 40, es de 19.092 observaciones, de las cuales 9.393 provienen del año 2004 y 9.699 del año 2024. Tras aplicar el proceso de limpieza y armonización de variables claves, se logró conservar 9.978 observaciones completas, lo cual representa un 52% del total.

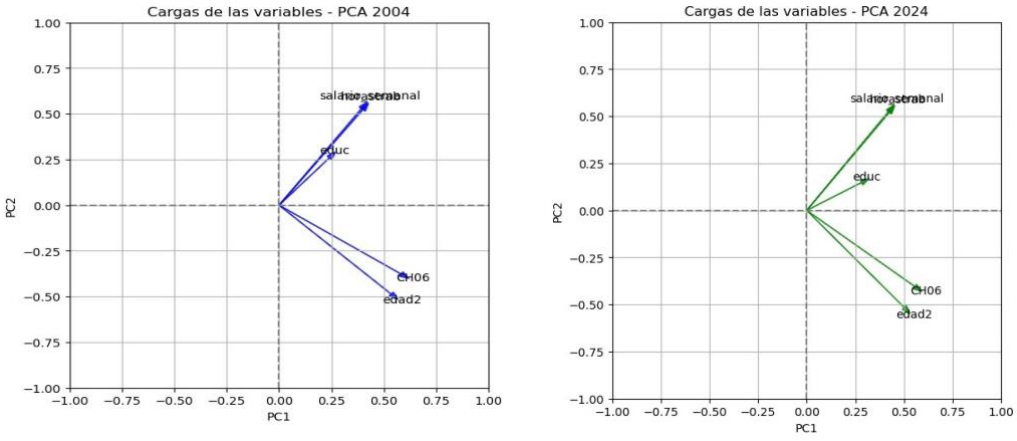




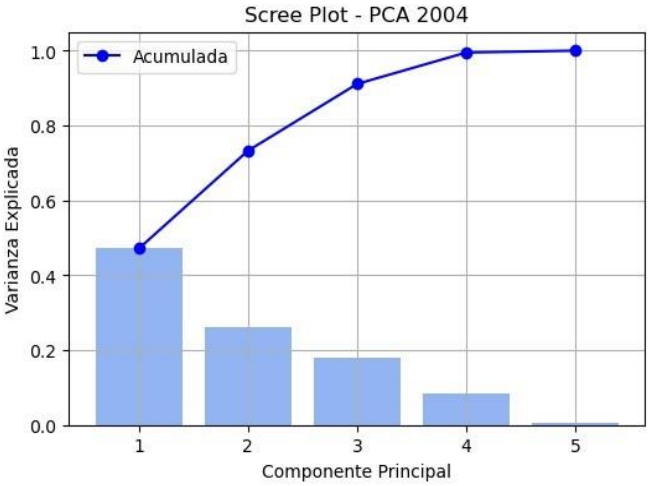
En la matriz de correlación del 2004: en salario semanal y horas trabajadas es 0.58, es una correlación moderada, a mayor cantidad de horas trabajadas, mayor salario semanal; en sexo (CH06) y edad es 0.96, es una correlación inusualmente alta. Esto sugiere que hay una fuerte diferencia etaria entre géneros en la muestra; en salario y educación es 0.19, es una correlación baja pero positiva. Sugiere que mayor educación se asocia con salarios algo más altos; en horas trabajadas y educación es 0.20, es una correlación similarmente baja. Sugiere que personas con más educación tienden a trabajar un poco más.
En la matriz de correlación del 2024: en salario semanal y horas trabajadas es 0.69, que subió respecto a 2004, mostrando una relación aún más fuerte entre ingresos y carga horaria; en salario y educación es 0.22, y horas trabajadas y educación es 0.23, en ambas correlaciones aumentaron levemente, lo que sugiere un ligero fortalecimiento del vínculo entre educación y desempeño laboral; en sexo y edad sigue siendo 0.96, lo que confirma que hay un patrón estructural o codificación fija en los datos que hace que estas dos variables estén fuertemente relacionadas.

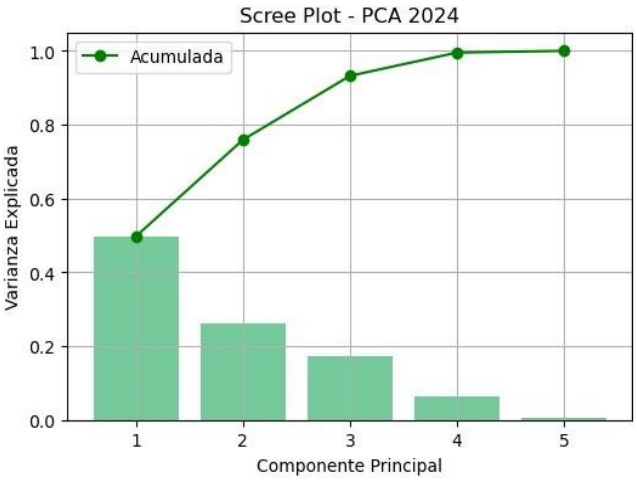


El PCA muestra que la estructura latente de los datos laborales presenta patrones similares entre 2004 y 2024, aunque hay indicios de mayor concentración y menor dispersión vertical en el año más reciente.

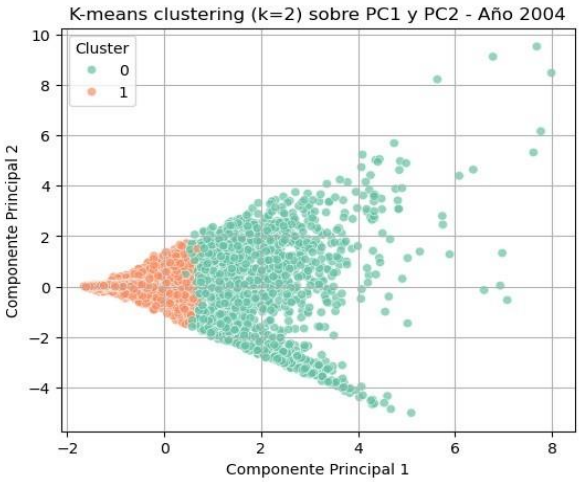


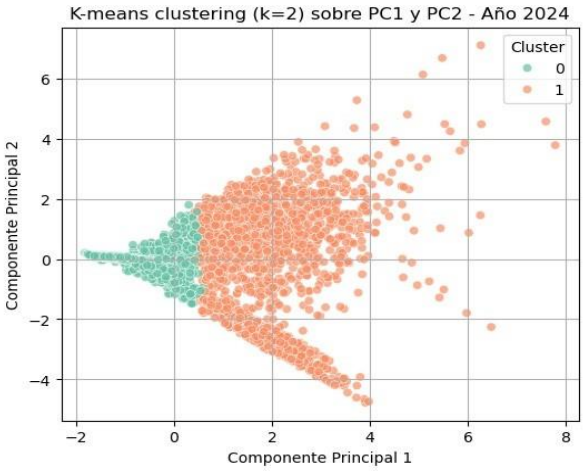
El perfil de cargas es bastante estable entre 2004 y 2024, lo que indica que las relaciones entre variables clave no cambiaron radicalmente. El educ tiene una carga positiva ligeramente más débil en 2024 que podría indicar una reducción en su capacidad explicativa principal, y CH06 y edad2 están más “aplastadas” hacia abajo en 2024, pero mantienen la misma dirección, reflejando consistencia estructural aunque con ligera variación en intensidad. PC1 sigue representando fuertemente la dimensión laboral (trabajo y salario), mientras que PC2 parece capturar diferencias demográficas, particularmente de edad y género. Y la superposición de salario y horas trabajadas refuerza su alta correlación y su rol como núcleo del primer componente.





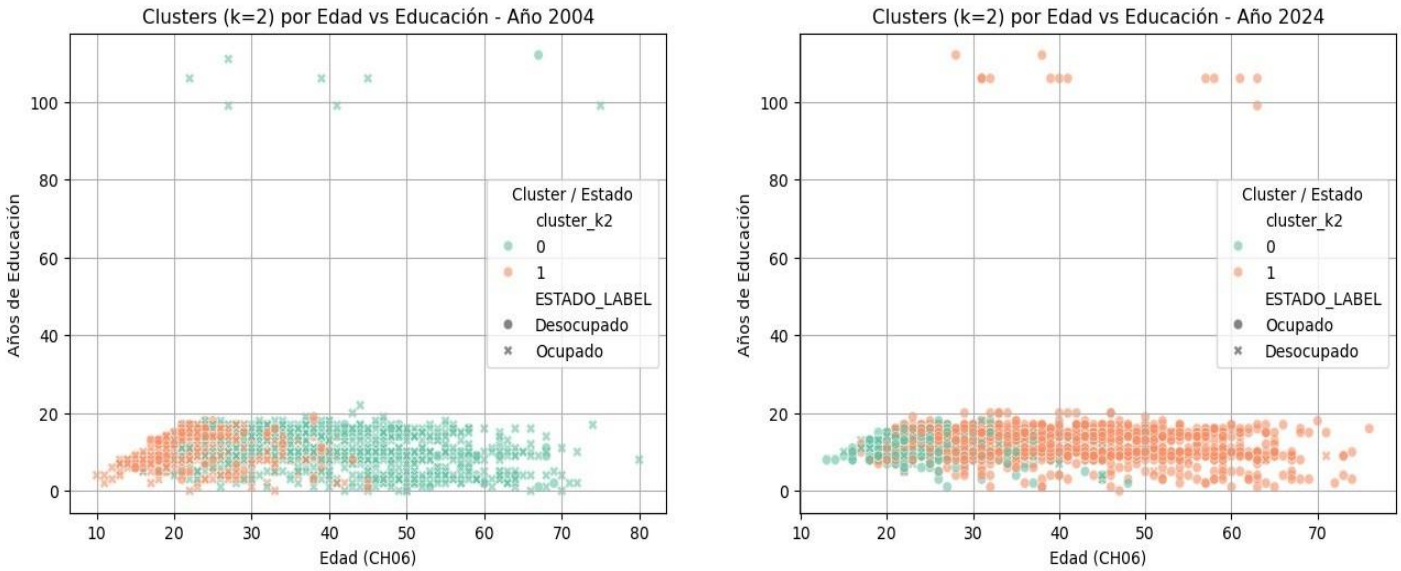
Los datos tienen una estructura similar en ambos años, con una clara concentración de información en los primeros dos componentes. El PC1 continúa dominando, principalmente vinculado a las variables laborales (horastrab y salario_semanal). Y la estructura demográfica (edad y género) sigue influyendo en el segundo componente (PC2), pero su peso relativo ha disminuido levemente.



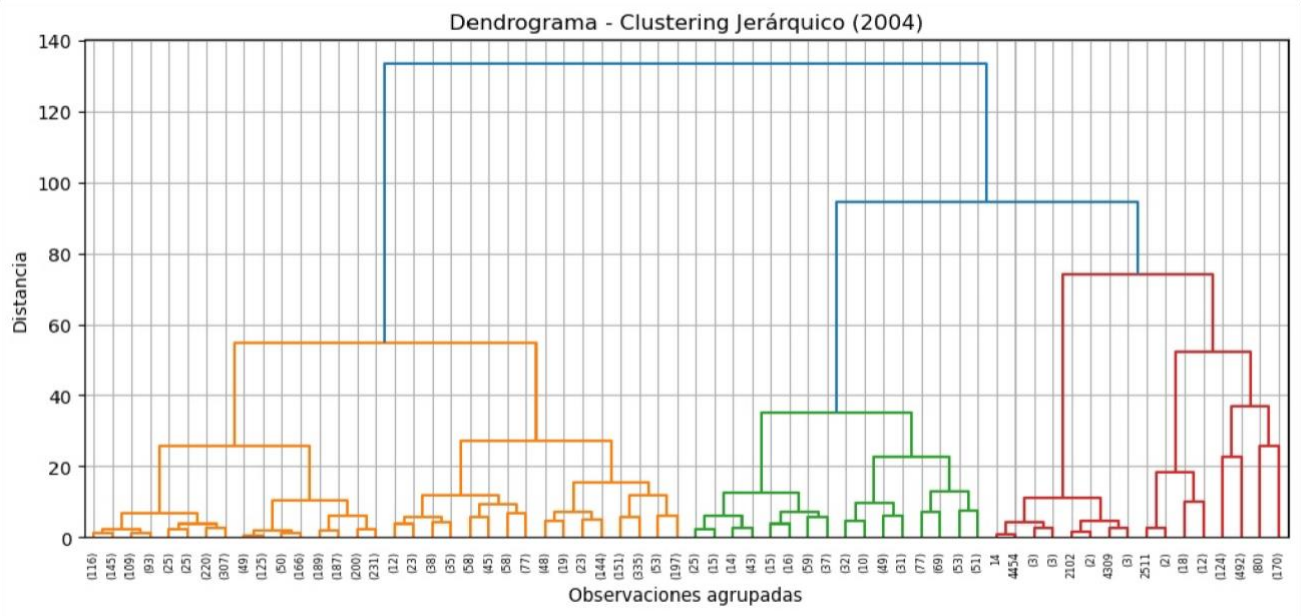


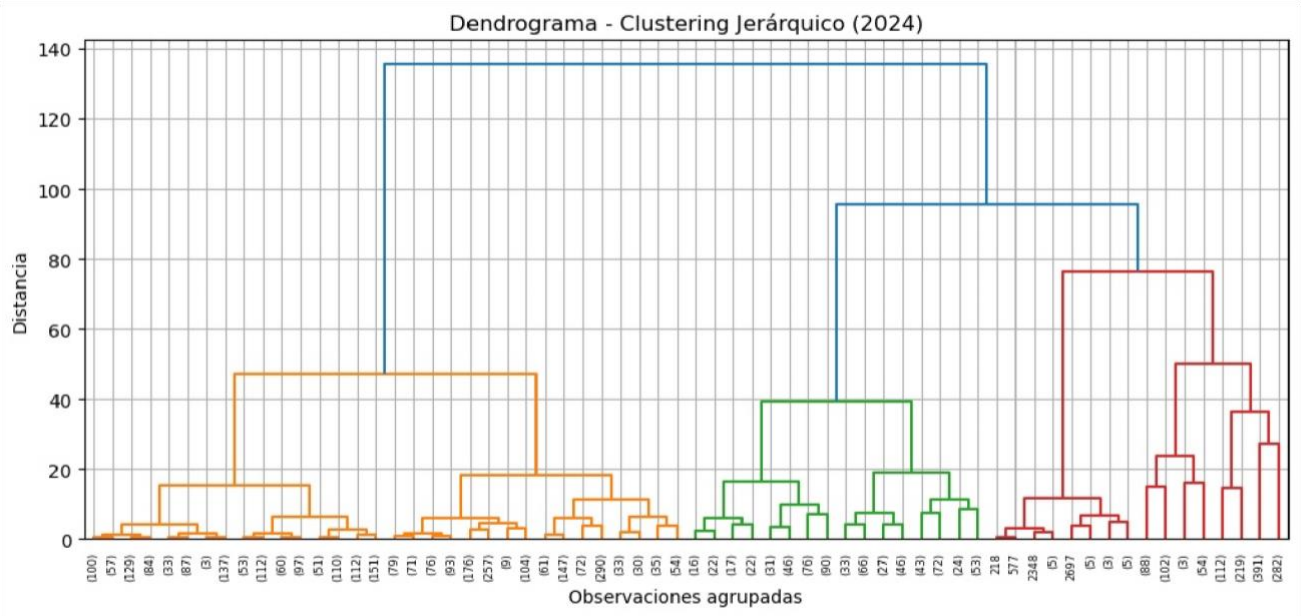
En 2004, el Cluster 0 (verde) agrupa a la mayoría, y el Cluster 1 (naranja) es más pequeño y concentrado alrededor del origen (valores bajos de PC1 y

PC2). Esto implica una estructura más homogénea, con un grupo pequeño que se distingue por menor desempeño o características.
En 2024, el Cluster 1 ahora domina claramente, ocupando casi toda la región, y el Cluster 0 se achica notablemente, limitado a la zona más densa y baja del gráfico. Esto sugiere una mayor polarización o diferenciación en la población, donde las características de “mayor PC1” se agrupan de forma más marcada.



En 2004 había una segmentación más clara entre jóvenes con baja educación (desocupados) y adultos con mayor educación (ocupados).
En 2024, esa frontera se difumina y el cluster más vulnerable crece. Esto podría reflejar un deterioro del rol de la educación como protector ante el desempleo o cambios estructurales en el mercado laboral.





En ambos años se mantienen estructuras jerárquicas con cuatro clústeres dominantes, pero en 2004, los grupos son más homogéneos entre sí, y en 2024, un grupo aparece claramente más distante, lo cual podría reflejar una segmentación más pronunciada en la población. Esto podría indicar un aumento de la polarización o fragmentación social en los datos del 2024 respecto al 2004.