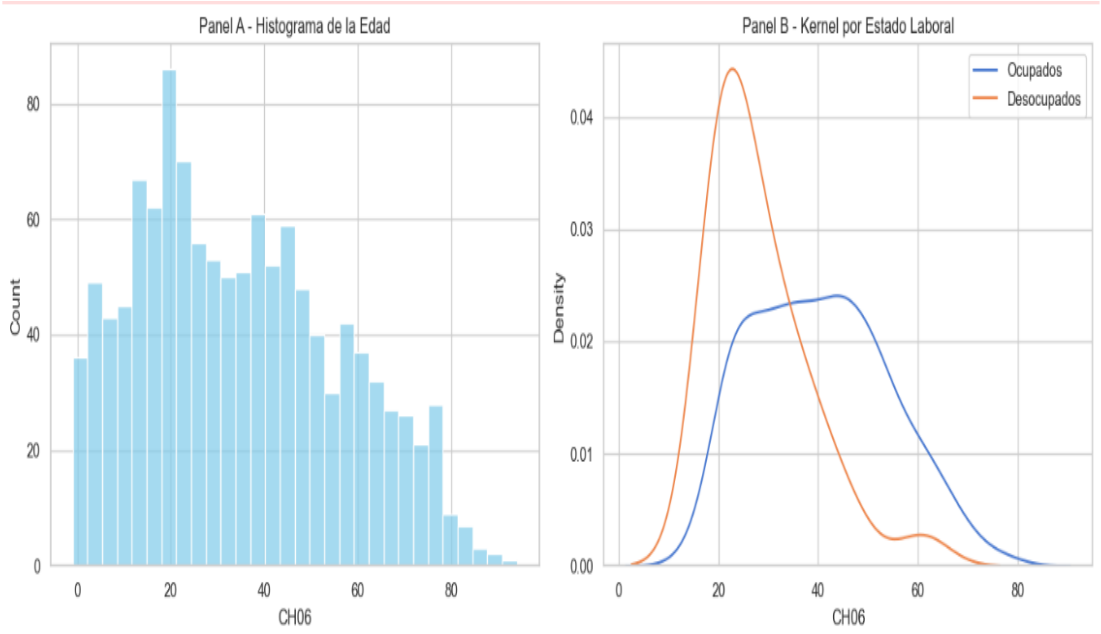
TRABAJO PRÁCTICO N° 3- Grupo 26

Parte I: Carga, limpieza y creación de variables

1) La distribución de edades en el Panel A muestra una concentración significativa en los grupos etarios más jóvenes, con un pico notable alrededor de los 15 a 20 años. A medida que aumenta la edad, la frecuencia disminuye progresivamente, lo cual es esperable en poblaciones donde hay más jóvenes que adultos mayores. Esta forma sesgada hacia la derecha indica una población relativamente joven. También se observa una caída notable en la cantidad de personas mayores de 60 años.

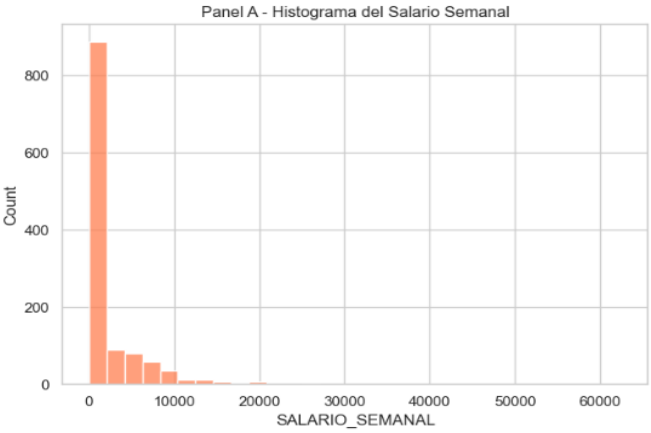


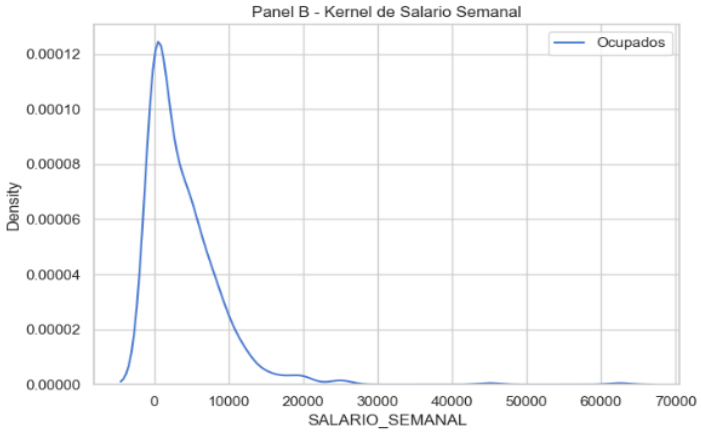

```
count    1193.000000
mean       16.805532
std        21.692615
min         0.000000
25%         0.000000
50%         0.000000
75%        36.000000
max        90.000000
Name: HORASTRAB, dtype: float64
       Observaciones  Con NaN en ESTADO  Ocupados  Desocupados  \
2004               0                  0         0            0
2024            1193                  0       576           29
Total           1193                  0       576           29

       Variables homogenizadas
2004                       178
2024                       178
Total                      178
```

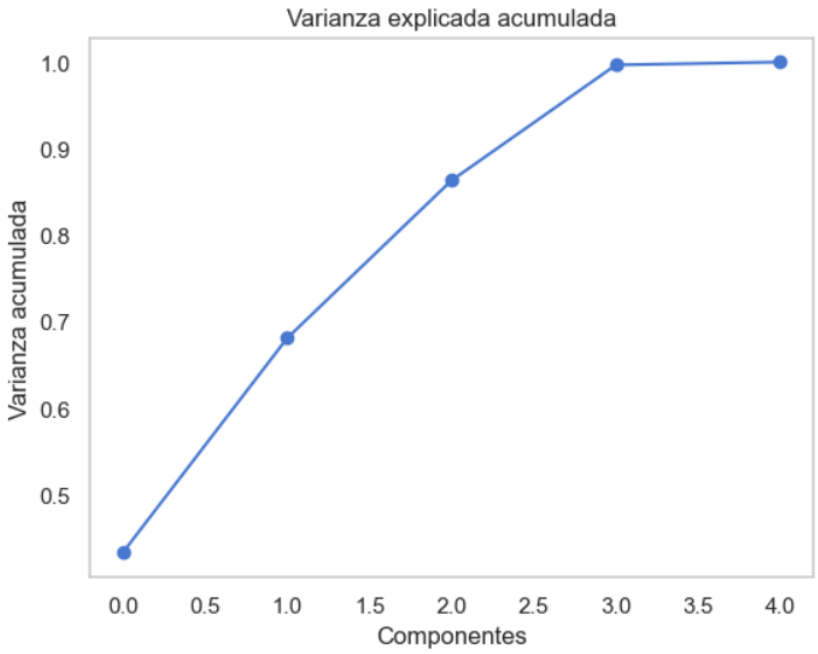
En el Panel A, el histograma muestra que la mayoría de los individuos tienen un salario semanal bajo, concentrándose fuertemente entre 0 y 10.000 pesos, con una clara asimetría positiva (cola larga a la derecha). En el Panel B, la distribución kernel para los ocupados refuerza esta observación: la densidad más alta se encuentra en los tramos bajos de salario, y luego disminuye rápidamente. Esto sugiere que los salarios están fuertemente concentrados en valores bajos, con pocos casos de ingresos muy altos. En conjunto, ambos gráficos evidencian una distribución desigual del salario entre los ocupados.

```
count    465.000000
mean       9.466667
std        7.532077
min        0.000000
25%        3.000000
50%       12.000000
75%       14.000000
max      111.000000
Name: EDUC, dtype: float64
```





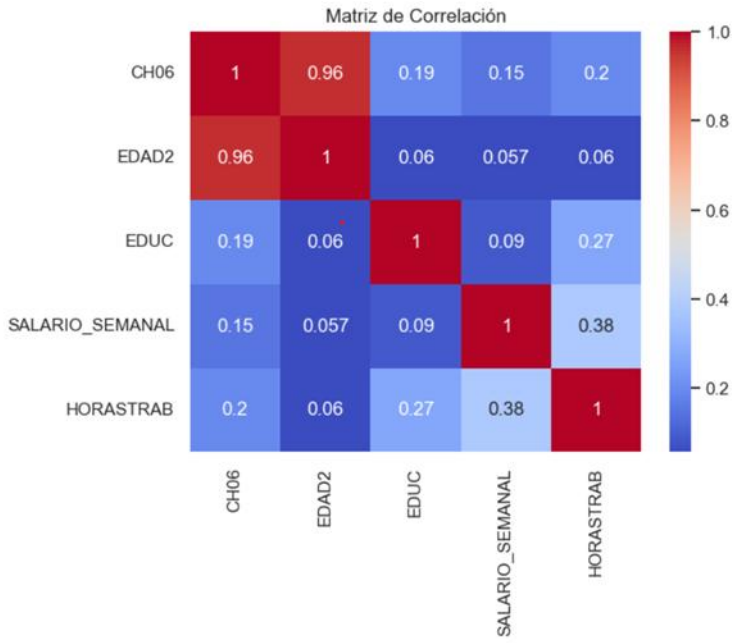
Cree la variable horastrab como el total de horas trabajadas como la suma de las horas en la ocupación principal y otras ocupaciones (PP3E_TOT+ PP3F_TOT). Presente una estadistica descriptiva (promedio, sd, min, p50, max) de dicha variable creada y comente



Parte II: Métodos no supervisados

Matriz de correlaciones con estos cinco predictores para su región y comente los resultados.

La matriz de correlación muestra una fuerte relación entre las variables CH06 (edad) y EDAD2, con un coeficiente de 0.96, lo cual indica que ambas variables capturan esencialmente la misma información. La correlación entre horas trabajadas y salario semanal es moderada (0.38), lo cual sugiere que trabajar más horas se asocia con un salario más alto, aunque no de forma perfecta. Por otro lado, la educación presenta una correlación baja tanto con el salario (0.09) como con las horas trabajadas (0.27), lo que podría indicar que el nivel educativo no está fuertemente relacionado con estas variables en esta muestra. En general, las correlaciones son bajas, lo que sugiere una relación débil entre la mayoría de las variables analizadas.

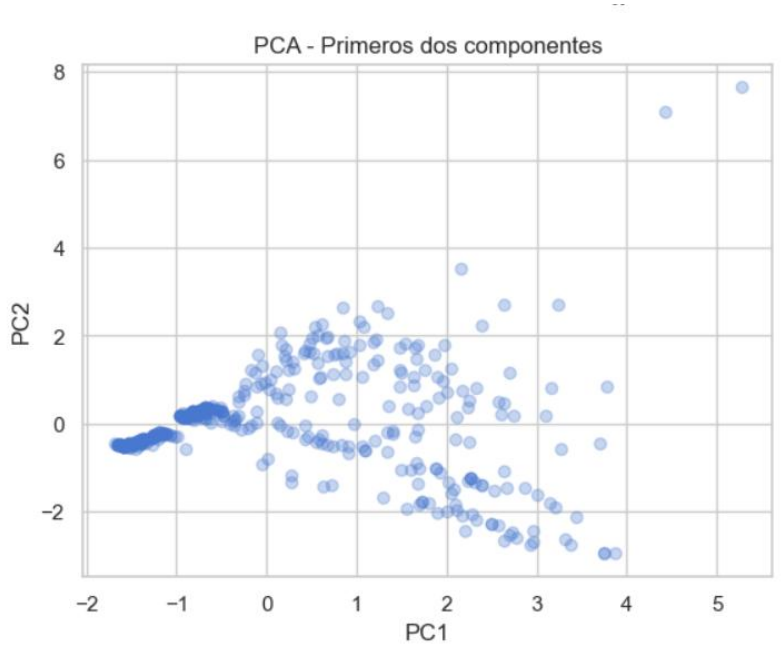


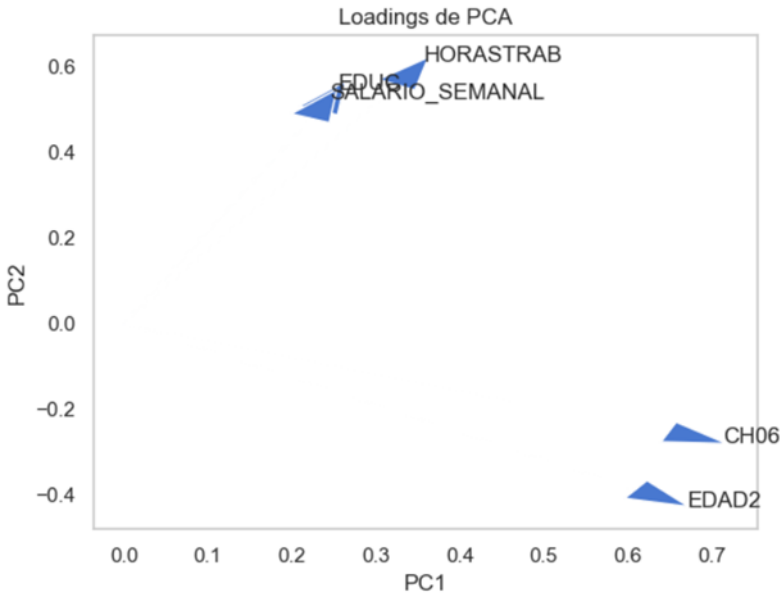
A.PCA:

PCA con salario: Apliquen PCA a las cinco variables seleccionadas para esta parte.

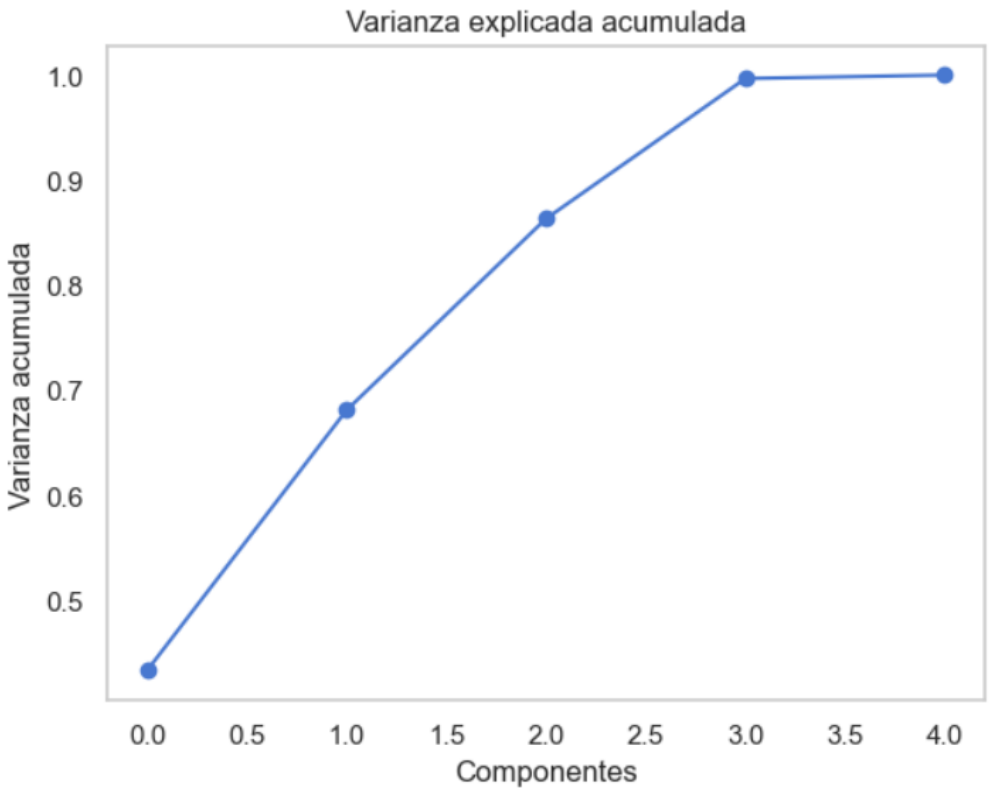
El gráfico muestra que la mayoría de las observaciones se agrupan cerca del origen, indicando que para muchas personas los valores combinados de las cinco variables estandarizadas no se desvían mucho del promedio. Sin embargo, hay observaciones que se dispersan especialmente hacia la derecha (PC1 alto) y hacia arriba (PC2 alto), lo que sugiere presencia de individuos con características atípicas (por ejemplo, altos salarios, mayor educación o más horas trabajadas). El hecho de que haya cierta forma de “abanico” sugiere correlaciones entre variables como salario y educación o edad y horas trabajadas.

Este análisis permite reducir la dimensionalidad del problema y detectar patrones o grupos con comportamientos distintos dentro de la población ocupada.





El gráfico muestra la varianza explicada acumulada por los componentes principales obtenidos mediante PCA. Podemos observar que los primeros dos componentes explican aproximadamente el 85% de la varianza total, lo cual indica que una gran parte de la información contenida en las cinco variables originales puede resumirse en solo dos componentes. Al incluir un tercer componente, se alcanza cerca del 99% de la varianza explicada, lo que sugiere que con tres componentes es posible capturar casi toda la variabilidad de los datos.



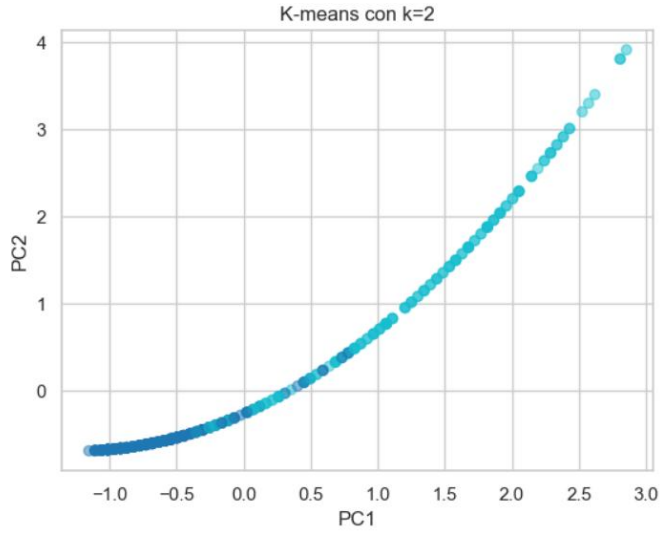
B.Cluster

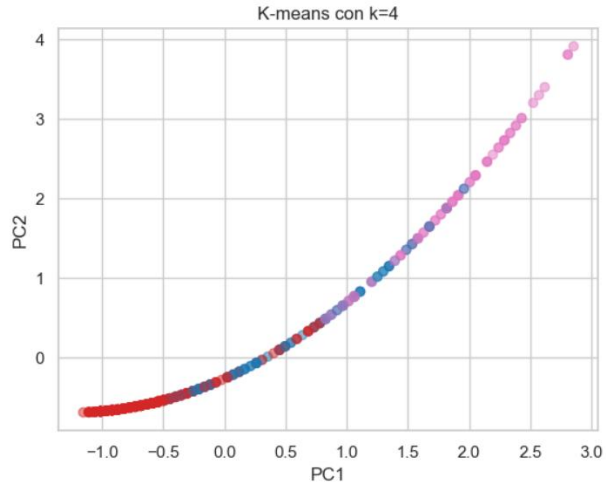
Cluster-K medias

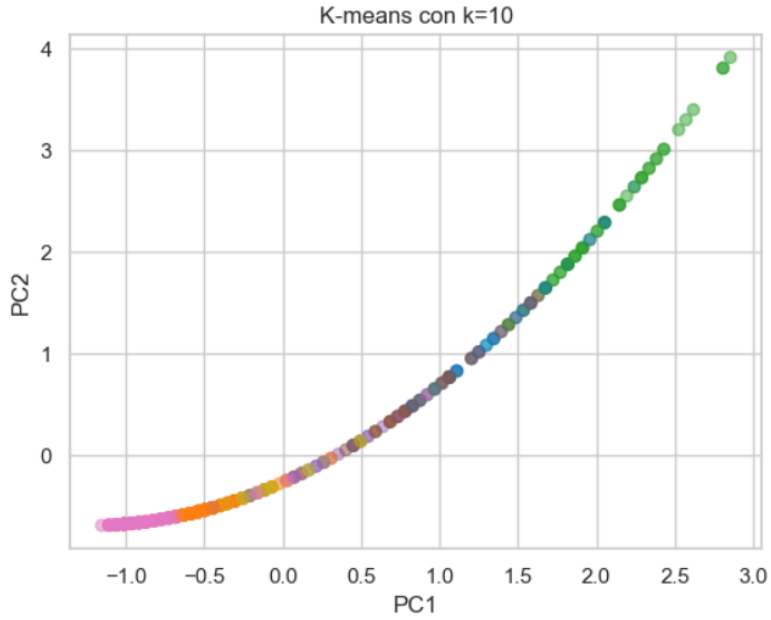
a.Corran el algoritmo con k=2, k=4 y k=10 usando n_init = 20, y grafiquen los resultados usando dos predictores. Interprétenlos.

En los gráficos con K-means para k=2, k=4 y k=10, se observa cómo varía la segmentación de los datos proyectados sobre los dos primeros componentes principales (PC1 y PC2).

- En k=2, los grupos son amplios y abarcan un rango grande de edades, lo que sugiere que hay una separación general entre dos perfiles socioeconómicos o demográficos.
- En k=4, los clusters comienzan a capturar mayor variación interna, probablemente diferenciando por rangos etarios más específicos o combinaciones con otras variables como ingresos o educación.
- En k=10, la segmentación es más detallada y granular. Esto podría reflejar distintos grupos etarios y niveles de ingresos o actividad, aunque el riesgo es que algunos clusters estén demasiado próximos entre sí o sean difíciles de interpretar.







Un dendograma es un diagrama en forma de árbol que se utiliza para representar visualmente los resultados de un análisis de agrupamiento

jerárquico (hierarchical clustering). Muestra cómo los datos se agrupan progresivamente en clústeres: al comienzo, cada observación es su propio grupo y, a medida que se sube en el árbol, los grupos se van fusionando según su similitud.En el eje horizontal se representan las observaciones o grupos, y en el eje vertical la distancia o disimilitud entre ellos. Cuanto más alta es la unión entre ramas, mayor es la diferencia entre los grupos que se están fusionando.