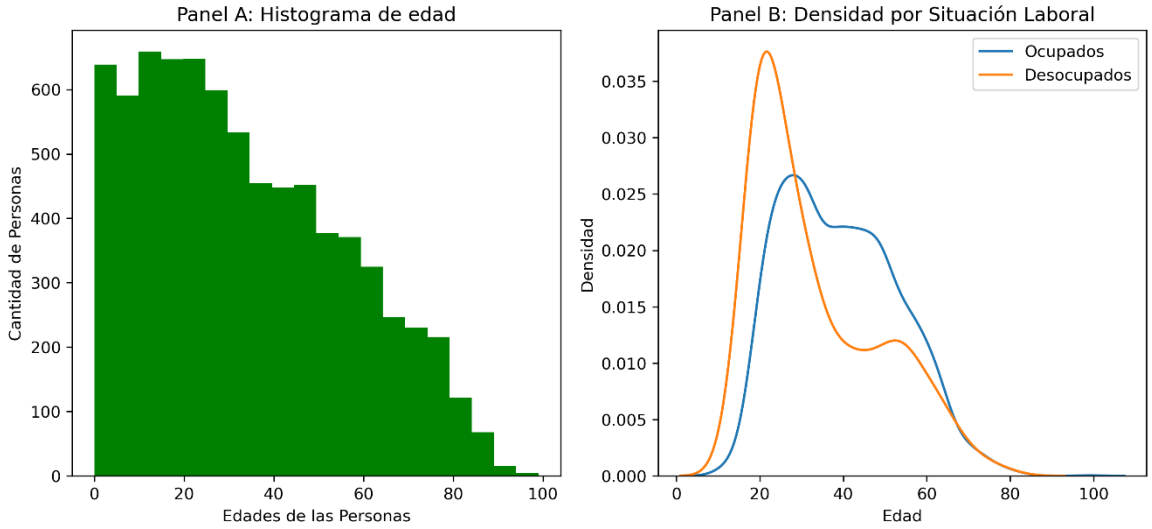
Trabajo Práctico N°3

Iván Martínez, Joaquín Reyes y Santiago Reinoso Sokol

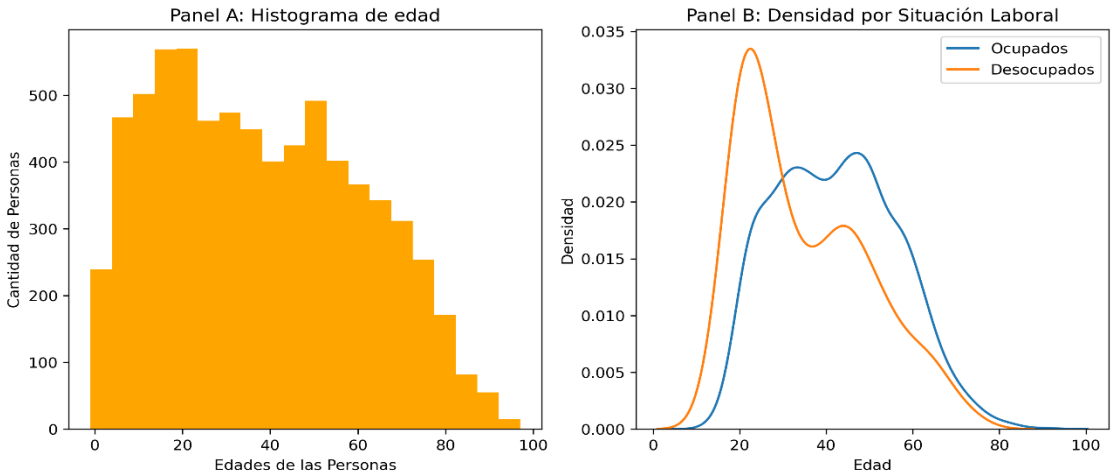
Parte 1

1) Gráfico 1 del año 2004



En el Gráfico 1 en su panel A podemos observar prácticamente que cuantos más años tienen, menos personas hay que contestaron, lo cual es lógico y esperable.
En el panel B podemos observar que hay un gran número de desocupados en personas de entre 15 a 25 años, lo cual desciende fuertemente hasta las personas de entre 45 a 55 años que hay un ligero rebote para luego seguir descendiendo y converger con la curva de ocupados. Esta última tiene un comportamiento menos fuerte comenzando un gran ascenso a partir de los 18 años para llegar a un punto máximo en torno a los 30 años, para luego ir descendiendo levemente hasta los 55 años y a partir de ahí una caída que converge con la de desocupados.

Gráfico 2 del año 2024



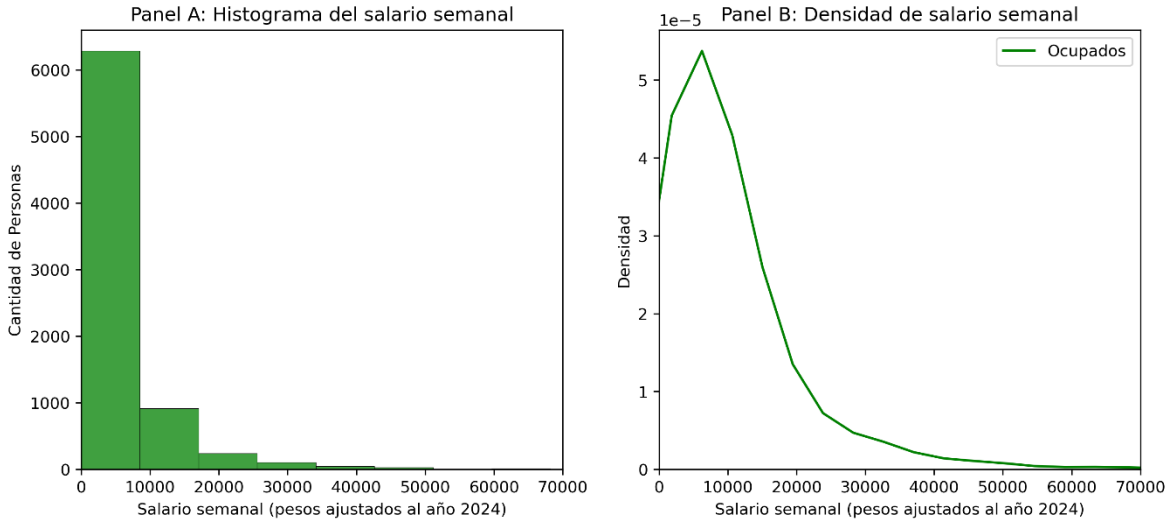
Mientras que en el Gráfico 2 en su panel A podemos observar que por personas de 0 a 10 años se respondió mucho menos que en el 2004 y además podemos ver como personas de entre 45 a 55 años respondieron mucho más, por eso aparece esa gran suba. Mientras que en el panel B pero del año 2024 podemos observar que la curva de desocupados se comporta de manera bastante similar a la curva del año 2004, teniendo una fuerte suba entre los 15 a 22 años alcanzando su punto máximo, para luego descender fuertemente hasta los 35 años donde hay un rebote hasta los 50 años donde vuelve a descender. Luego la curva de ocupados se comporta similar a la curva de ocupados del año 2004 pero esta tiene un punto máximo entre los 45 a 55 años para luego comenzar a descender.

2) Tabla de 2004 y 2024

mean	**7.97**
std	**7.94**
min	**0.0**
p50	**7.0**
max	**99.0**

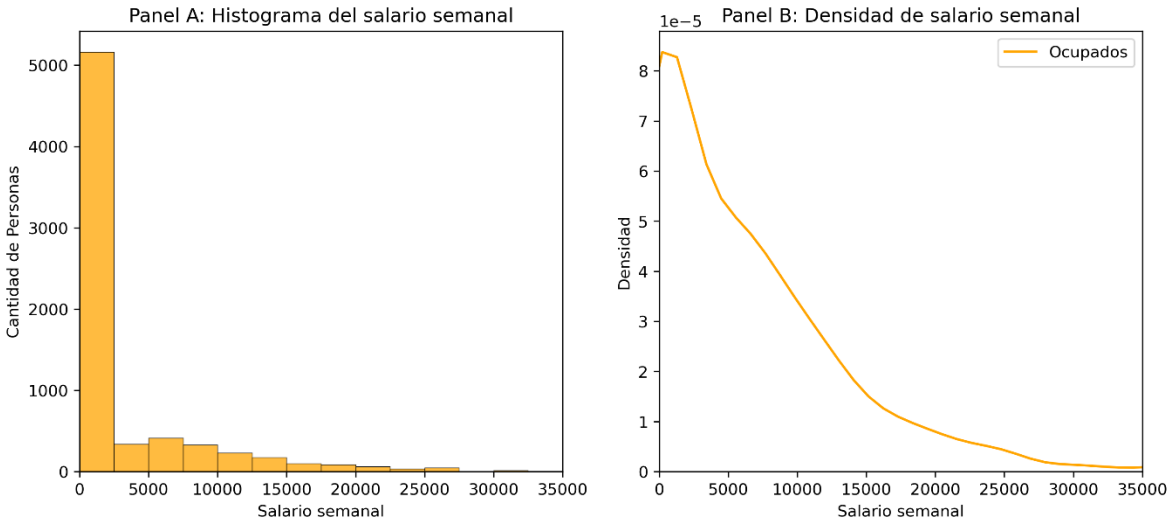
La media de años de educación formal del año 2004 y 2024 es de casi 8 años aproximadamente, con una mediana de 7 años. El mínimo es 0, lo que indica personas sin educación formal, en el máximo tuvimos un error que no pudimos resolver o encontrar el error en el Código, se toma el 99 de ns/nr como respuesta (hay pocas observaciones de esto igualmente). La distribución muestra una gran dispersión, de aproximadamente 8 años, reflejando diferentes niveles alcanzados por la población.

3) Gráfico 3 del año 2004



En el Panel A del Gráfico 3 observamos que el promedio de las personas tiene un salario semanal de aproximadamente $4150, siendo convertidos a valores del año 2024, lo que indica una alta concentración en salarios muy bajos, viéndose sesgado por la gente que no responde de manera cierta. Con un monto máximo de $852525 lo cual provoca un aumento en el promedio. En el Panel B, la distribución de ocupados es prácticamente idéntica al histograma de salario semanal, lo que quiere decir que los desocupados no respondieron.

Gráfico 4 del año 2024



Mientras que en el Gráfico 4 de su Panel A se puede observar que el promedio de las personas tiene un salario semanal de aproximadamente $3010, lo que indica una alta concentración en salarios muy bajos, viéndose sesgado por la gente que no responde de manera cierta. Con un monto máximo de $200000. En el Panel B, la distribución de ocupados es prácticamente idéntica al histograma de salario semanal, lo que quiere decir que los desocupados no respondieron.

4) Tabla 2

mean	**18.27**
std	**58.08**
min	**0.0**
p50	**0.0**
max	**1044.0**

En la Tabla 2 la cantidad de horas trabajadas por semana del año 2004, siendo la suma de las horas de la ocupación principal y otras ocupaciones en promedio son un poco más de 18hs, con una dispersión muy alta, siendo aproximadamente 58 hs y un máximo de 1044hs lo cual eleva el promedio, siendo una cantidad imposible.

Tabla 3

mean	**19.96**
std	**71.84**
min	**0.0**
p50	**0.0**
max	**1998.0**

Mientras que en la Tabla 3 la cantidad de horas trabajadas por semana del año 2024, siendo la suma de las horas de la ocupación principal y otras ocupaciones en promedio son casi

20hs, con una dispersión más alta que en 2004, siendo casi 72 hs y un máximo superior de 1998hs lo cual eleva el promedio por mucho, siendo imposible ese número.

5) Tabla 4

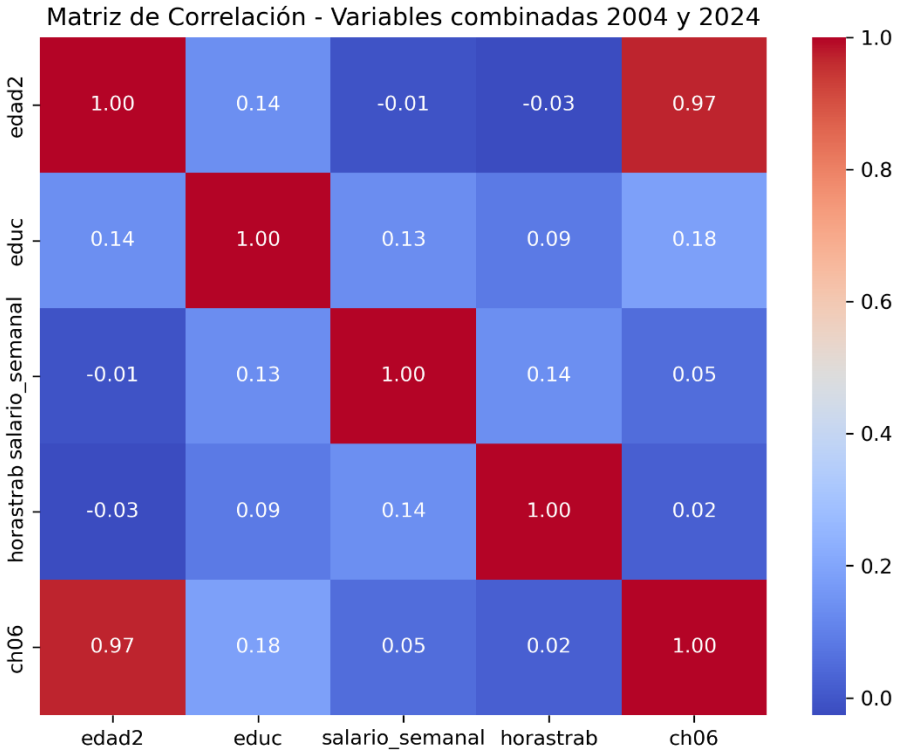
Tabla 1. Resumen de la base final para la region YYY

	2004	2024	Total
Cantidad observaciones	7.647	7.051	14.698
Cantidad de observaciones con Nas en la variable "Estado"	0	0	0
Cantidad de Ocupados	3.079	3.224	6.303
Cantidad de Desocupados	528	311	839
Cantidad de variables limpias y homogeneizadas	8	8	16

Durante el proceso de limpieza de datos, se identificaron algunas variables que contenían valores nulos, ceros o la ausencia de información. En primer lugar, se eliminaron las filas que contenían valores nulos en variables clave, ya que su presencia podría sesgar las estadísticas o afectar la interpretación. En algunos casos, se aplicaron filtros adicionales para identificar valores que aparecían como cadenas de texto con ceros ("0.0") se transformaron y luego se eliminaron. Esta limpieza fue fundamental para garantizar que las estadísticas descriptivas y los gráficos reflejaran las características reales de la población del Gran Buenos Aires. Así mismo llevamos a cabo la homogeneización de diversas variables de interés ya que en su mayoría las variables del año 2024 eran simplemente números y las variables del año 2004 eran strings que decían la respuesta, con eso logramos que las variables de ambos años se expresen por igual en forma de strings.

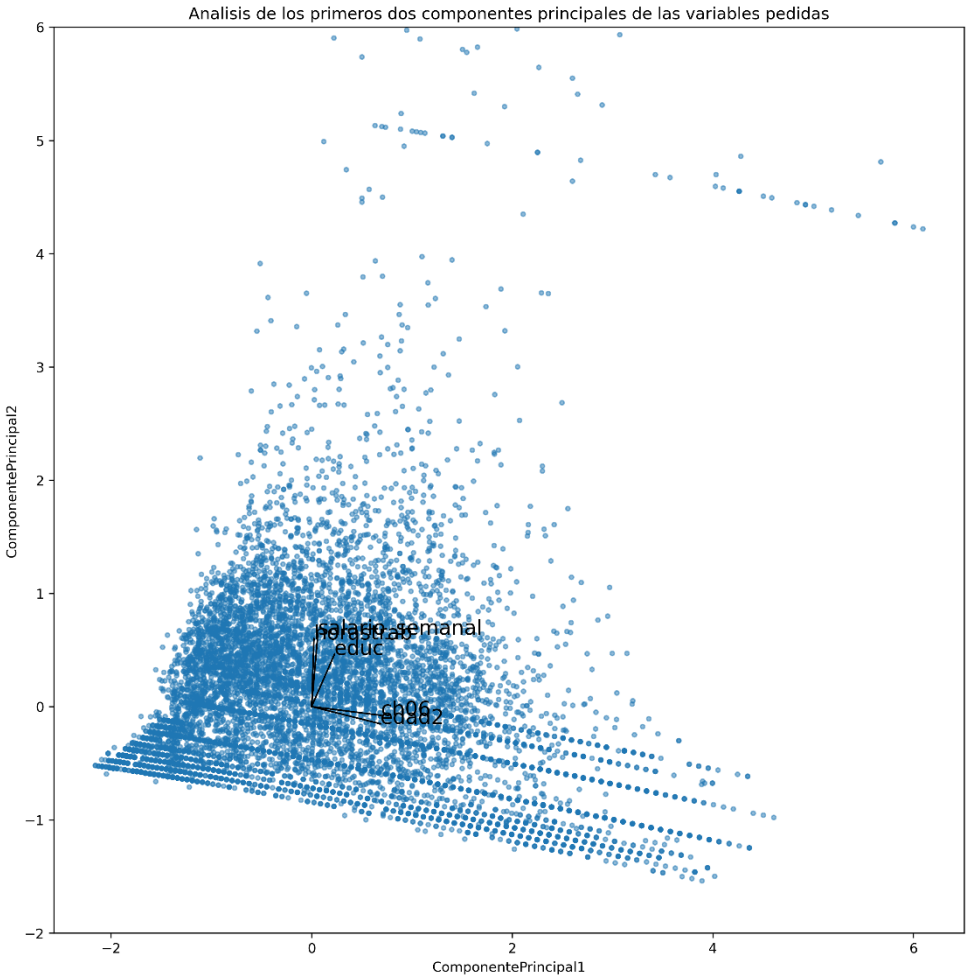
PARTE 2

1)



En esta matriz podemos observar que la relación entre variables combinadas de 2004 y 2024, en este caso las 4 variables creadas anteriormente y **ch06** (edad). Se puede ver por ej. Que hay una relación directa muy fuerte entre **edad2** y **edad** (tiene sentido ya que **edad2** es la **edad** al cuadrado) mientras que podemos ver otras relaciones no tan potentes pero visibles como la relación entre **educ** y **ch06** con una correlación de 0,18 o la correlación entre **horastrab** y **salario_semanal** con un coeficiente de correlación de 0,14

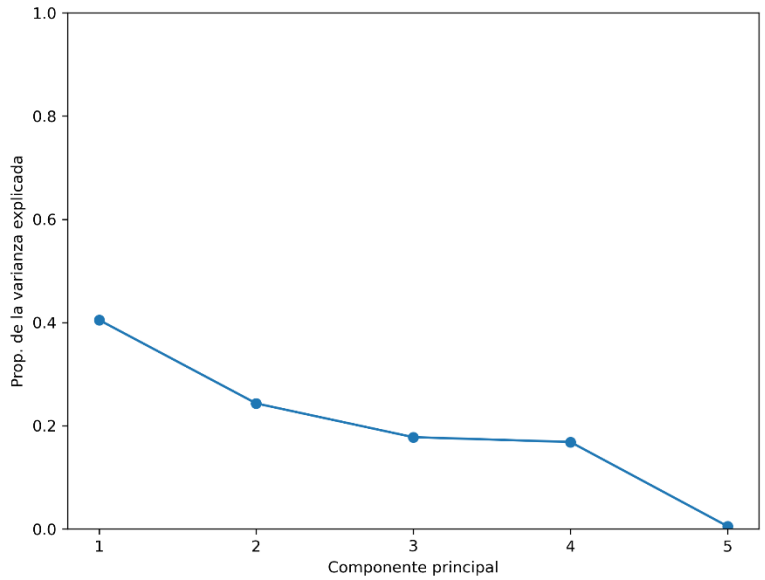
2) Y 3)



2. Con este grafico podemos ver las relaciones entre los componentes principales del PCA, donde se ve una pequeña relación negativa, que se puede deber a que edad y edad2 tienen datos parecidos que siguen una misma dirección, pudiendo verse como una agrupación de estos datos y toda la dispersión de la izquierda puede deberse a las otras variables que ponen peso llevando hacia arriba los scores.

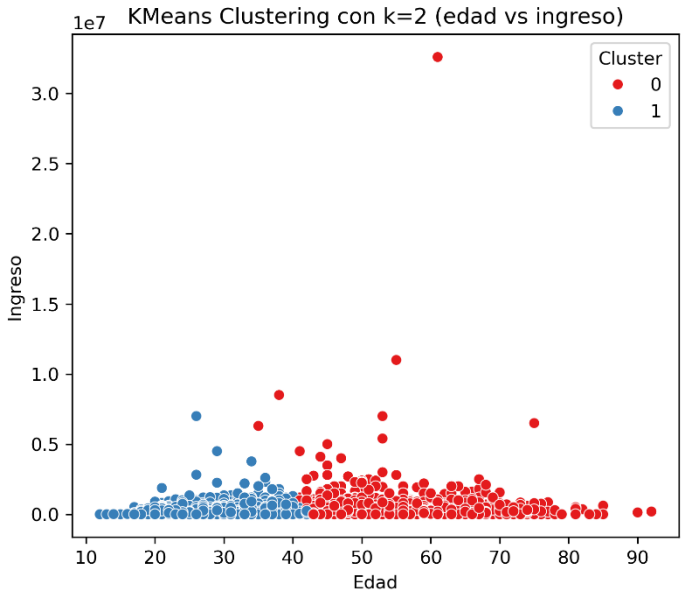
3) Con los loading podemos ver que claramente la edad y edad2 llevan un peso hacia abajo mientras que educ, salario_semanal y horastrab ponen el peso de los scores hacia arriba. Y se ve que las variables con más peso son el salario_semanal, edad2 y ch06, mientras que educ y horastrab tienen un peso un poco menor.

4)

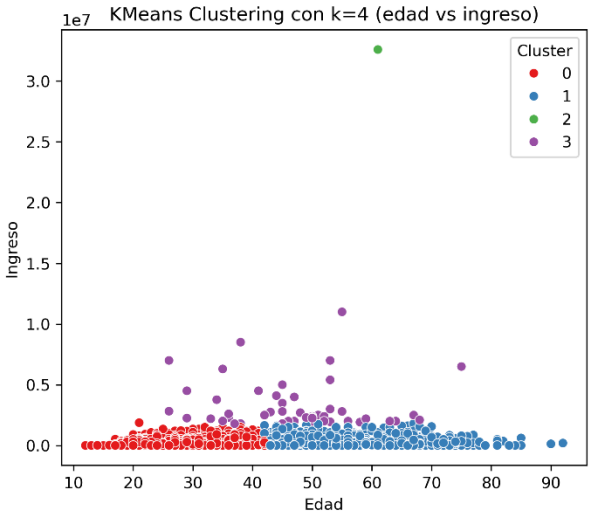


En este grafico podemos ver como se comparan las diferentes varianzas explicadas de los 5 componentes principales del pca realizado, donde podemos ver una varianza explicada del 40% aprox en el componente principal 1, bajando a 25% en el componente principal 2, a 20% en el componente principal 3, a 18% en el componente principal 4 y a casi 0% en el componente principal 5

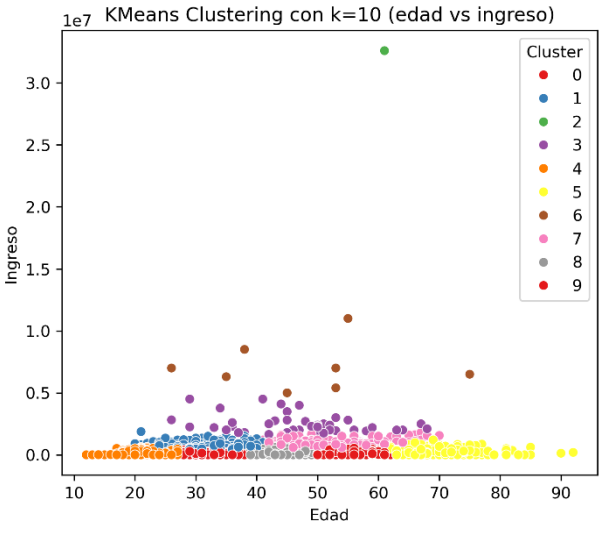
5) a) A partir de los resultados podemos ver que:
Al separarse en 2 Clusters se generan grupos donde uno tiene más observaciones, pero aun así la diferencia no es tan amplia y vemos que los grupos de forman de modo tal que las menores edades corresponden a un grupo y las mayores a otro, pero no se puede decir lo mismo en cuanto a los ingresos, ya que esta variable está distribuida de forma similar en ambos grupos. Entonces podemos decir que K=2 es especialmente útil para dividir en cuanto a una variable (edad en este caso), es decir, simplemente se da una división entre jóvenes y adultos mayores.



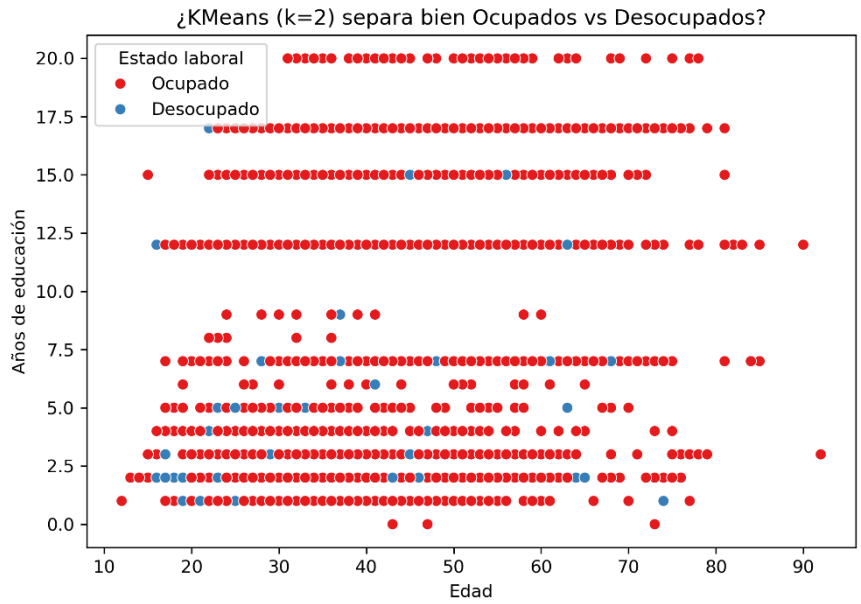
Al separarse en 4 clusters podemos ver la misma división que antes en cuanto a las edades, pero ahora también hay una división visible en los ingresos, ya que las observaciones correspondientes a mayores ingresos corresponden a su propio Cluster, mientras que también se divide de forma apropiada en cuanto a la edad. Por último, hay una observación muy alejada del resto ya que su ingreso es muy superior a las otras, entonces esta observación es la única de su propio Cluster ya que de ser introducida en otro grupo la varianza con respecto a otras observaciones sería muy grande.



Al separarse en 10 clusters podemos ver una mayor segmentación, ya que ahora hay grupos de jóvenes con ingresos altos, jóvenes con ingresos bajos, adultos con ingresos altos, adultos con ingresos altos y demás. Esto permite disminuir la varianza de los clusters ya que ahora son más homogéneos y la mayor segmentación ayuda a comprender mejor las observaciones.



b) Podemos ver cómo se diferencian claramente los ocupados de los desocupados.



6) Un dendograma es un diagrama utilizado para observar cómo está agrupado un conjunto de datos. Se dice que este tiene forma de árbol del cual se desprenden “ramas”. Este tipo de diagrama es útil para ver de donde provienen los datos, si hay subgrupos, como se organizan y demás.