<https://colab.research.google.com/drive/1GM0I3FaNiNZYxXWvXUYVZo8BqpuSg6Tu#scrollTo=gZaYPrHfWsnB>

### **Introducción**

En el competitivo mercado actual, entender a tus clientes es crucial para el éxito de cualquier negocio. La segmentación de clientes, que consiste en dividir un mercado en grupos distintos de clientes que comparten características similares, es una herramienta poderosa que permite a las empresas identificar y satisfacer las necesidades únicas de diferentes grupos de clientes. Al aprovechar la segmentación de clientes, las empresas pueden superar a la competencia desarrollando productos y servicios que atraigan de manera única a cada segmento.

### **Resumen del Proyecto**

Nuestro proyecto se centra en la segmentación de clientes utilizando un conjunto de datos que contiene información básica sobre los clientes, incluyendo su ID, sexo, estado civil, edad, educación, ingresos, ocupación y tamaño del asentamiento. El objetivo es analizar estos datos para identificar segmentos de clientes distintos dentro de nuestro supermercado. Al entender estos segmentos, podemos proporcionar información práctica a nuestro equipo de marketing para desarrollar estrategias dirigidas que resuenen con cada grupo.

### **¿Por qué la Segmentación de Clientes?**

La segmentación de clientes permite a las empresas:

1. **Identificar Clientes Objetivo**: Entender quiénes son los clientes más valiosos, qué necesitan y cómo atenderlos mejor.
2. **Mejorar la Satisfacción del Cliente**: Adaptar productos y servicios para satisfacer las necesidades específicas de diferentes segmentos, aumentando así la satisfacción y lealtad del cliente.
3. **Optimizar Estrategias de Marketing**: Desarrollar campañas de marketing más efectivas dirigidas a grupos específicos de clientes con mensajes personalizados.
4. **Incrementar los ingresos**: Enfocarse en clientes y oportunidades de alto valor, lo que lleva a mayores tasas de conversión y aumento de ventas.

### **Aplicación en el Mundo Real**

Imagina un supermercado que utiliza tarjetas de membresía para recopilar datos básicos sobre sus compradores. Con estos datos, el supermercado puede realizar una segmentación de clientes para obtener información sobre quiénes son sus clientes. Por ejemplo, podrían descubrir que los jóvenes profesionales con altos ingresos prefieren comprar productos orgánicos premium, mientras que las familias con ingresos moderados están más interesadas en ofertas de valor. Con este conocimiento, el supermercado puede adaptar su oferta de productos, diseño de la tienda, promociones y esfuerzos de marketing para atender a estos segmentos distintos, mejorando así la experiencia de compra y aumentando las ventas.

### **Propuesta de Inversión**

Buscamos inversión para avanzar en nuestro modelo de segmentación de clientes, lo que nos permitirá ofrecer información precisa y práctica a nuestro equipo de marketing. Esta inversión se utilizará para:

1. **Mejorar la Recopilación de Datos**: Mejorar la calidad y cantidad de datos de clientes recopilados a través de tarjetas de membresía y otros canales.
2. **Desarrollar Analíticas Avanzadas**: Construir modelos de aprendizaje automático sofisticados para analizar datos de clientes e identificar segmentos significativos.
3. **Implementar Estrategias Dirigidas**: Desarrollar y ejecutar estrategias de marketing adaptadas a cada segmento de clientes para impulsar las ventas y aumentar la lealtad del cliente.
4. **Escalar Operaciones**: Expandir la aplicación del modelo a múltiples ubicaciones para obtener una comprensión integral de nuestra base de clientes.

Al invertir en este proyecto, nos ayudarás a revolucionar la forma en que entendemos e interactuamos con nuestros clientes, lo que llevará a una experiencia de compra más personalizada, una mayor satisfacción del cliente y, en última instancia, una mayor rentabilidad para nuestro supermercado.

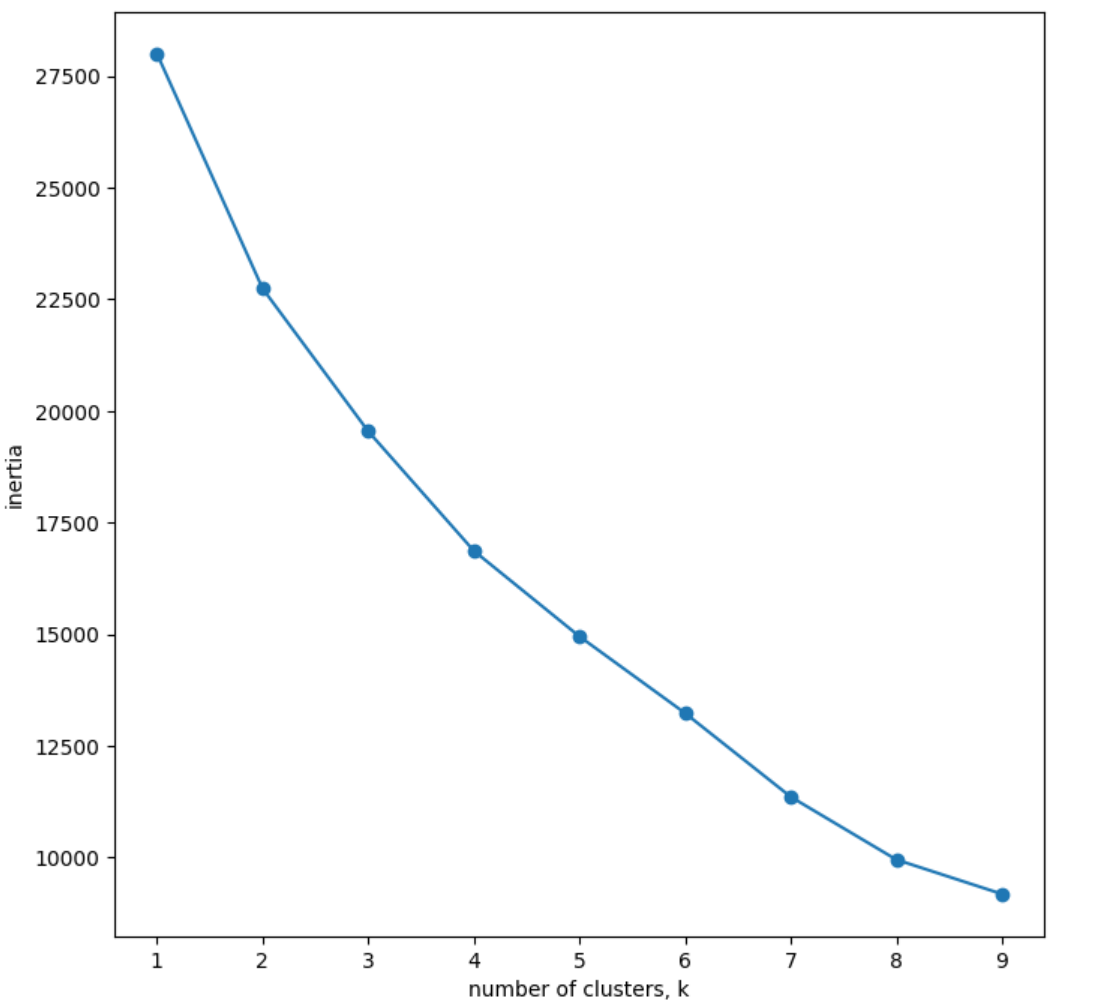
### **Proyecto**

#### **Preprocesamiento de Datos**

Se realizan reemplazos de valores en las columnas Education, Occupation, y Settlement size para convertirlas a etiquetas más legibles. Luego, se convierten estas columnas categóricas en variables dummy utilizando pd.get\_dummies.

#### **Aplicación de K-Means**

Se utiliza el algoritmo de clústering K-Means para segmentar los datos. Se elige el número óptimo de clústers mediante la evaluación de la inercia para diferentes valores de k.

Para elegir el valor óptimo de K usamos el método del codo. Es un método gráfico para encontrar el valor K óptimo en un algoritmo de agrupamiento de k-means. El gráfico del codo muestra los valores de suma de cuadrados dentro del grupo (WCSS) en el eje y correspondientes a los diferentes valores de K (en el eje x). El valor óptimo de K es el punto en el que el gráfico forma un codo.

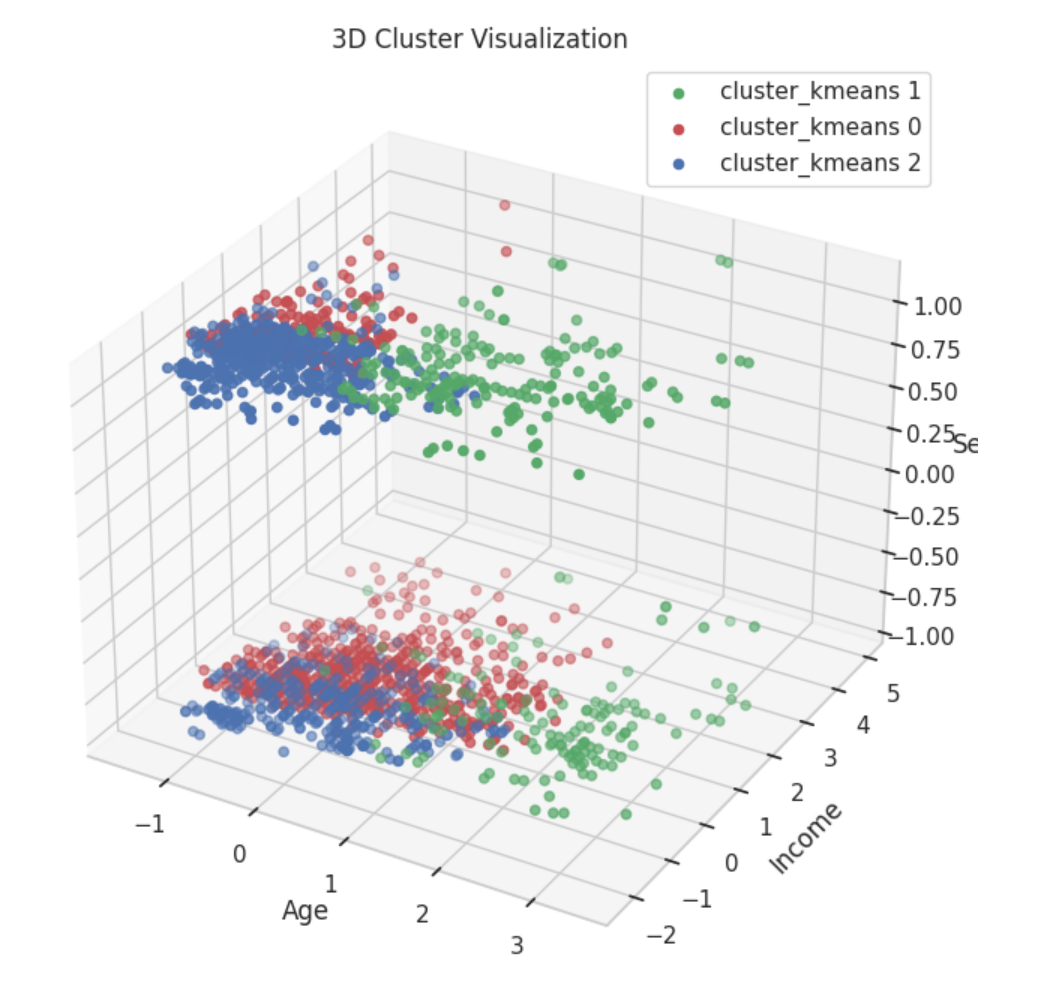
No parece tan claro el codo como para poder determinar la cantidad óptima de clústers. Vamos a probar con 3, 4 y 5.

Los ejes que elegiste para graficar en el análisis de segmentación de clientes son Age (Edad), Income (Ingreso) y Sex (Sexo). Estas variables se consideran importantes para segmentar clientes por las siguientes razones:

La edad de los clientes es fundamental para comprender sus necesidades y comportamientos de compra. Por lo general, diferentes grupos de edad tienen preferencias y hábitos de consumo distintos.

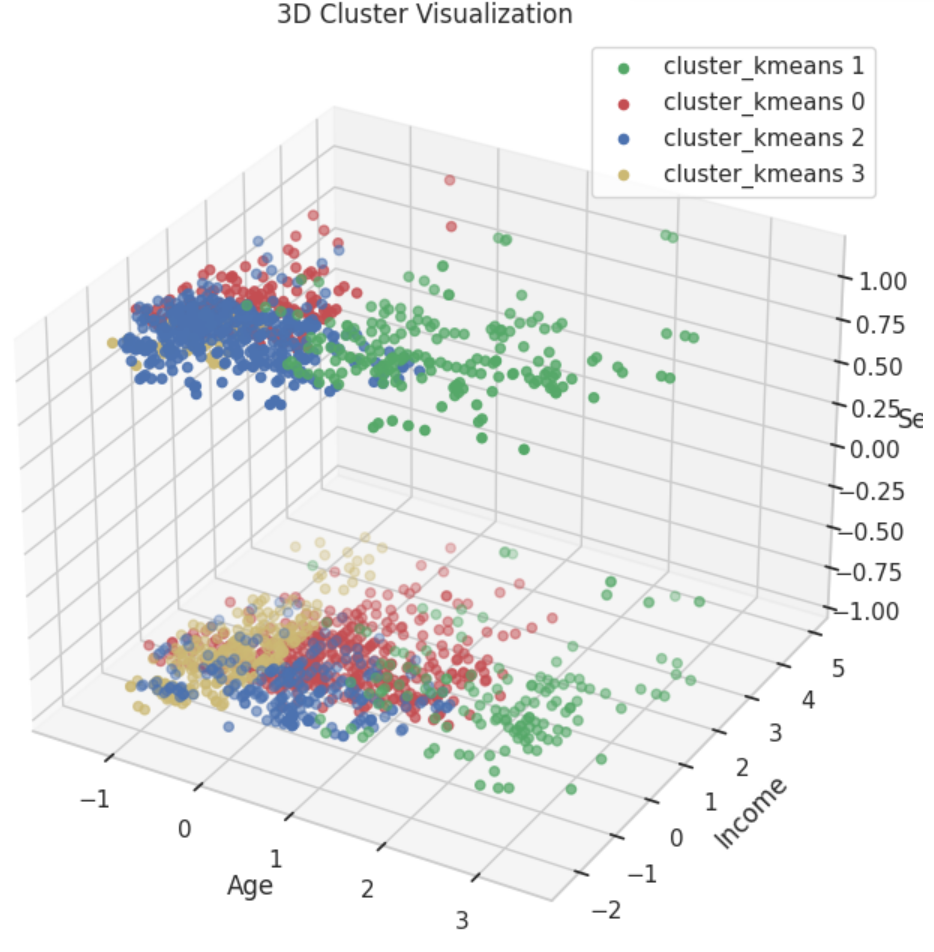
El ingreso económico de los clientes es un indicador clave de su capacidad de gasto y poder adquisitivo. Clientes con diferentes niveles de ingresos probablemente tendrán preferencias de compra diferentes.

El sexo de los clientes también puede influir en sus decisiones de compra y preferencias. Algunos productos o servicios pueden ser más populares entre un género específico.



Viendo solo estos ejes podemos ver:

* clúster 0: mayormente formada por hombres jóvenes de ingresos medio-altos.
* clúster 1: gente mayor de ambos géneros y de ninguna clase económica en particular.
* clúster 2: mayormente mujeres jóvenes de ingresos medio-bajos.

Probando con 4 clústers:

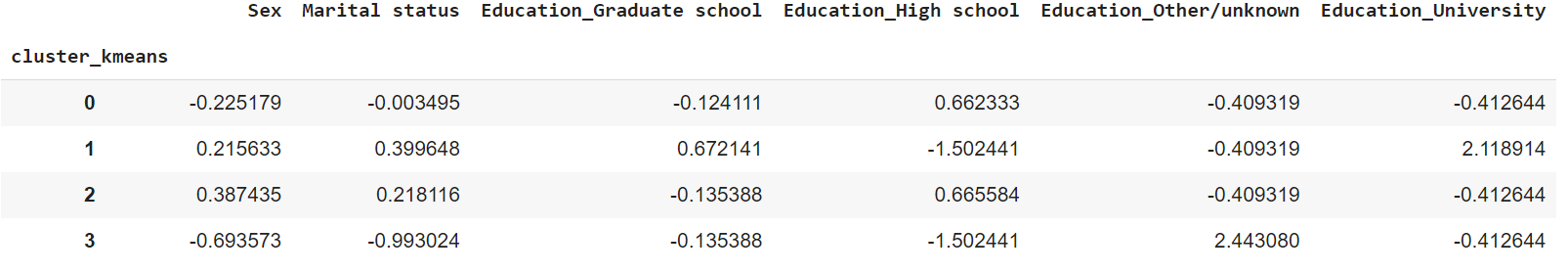
Lo que se ve a simple vista:

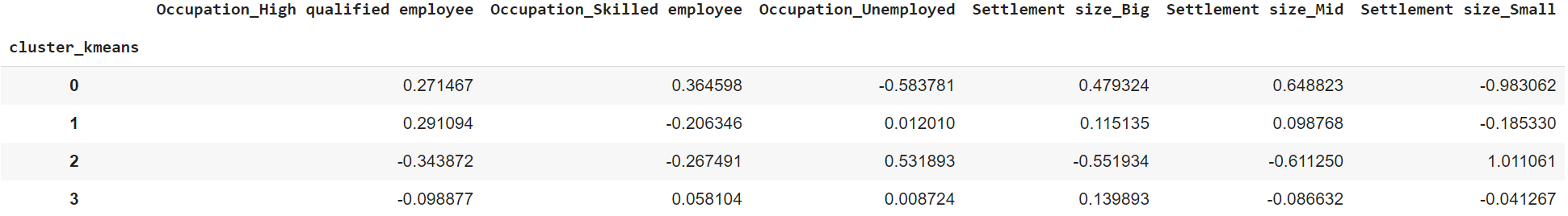
* clúster 0: Es similar al clúster 0 (hombres jóvenes) del gráfico anterior.
* clúster 1: al igual que el clúster 2 del gráfico anterior, gente mayor.
* clúster 2: muy parecido al clúster 1, mujeres jóvenes de clase media-baja
* clúster 3: hombres muy jóvenes.

Parece que clasifica mejor a los hombres que conformaban el clúster 0 en clústers 0 y 3.

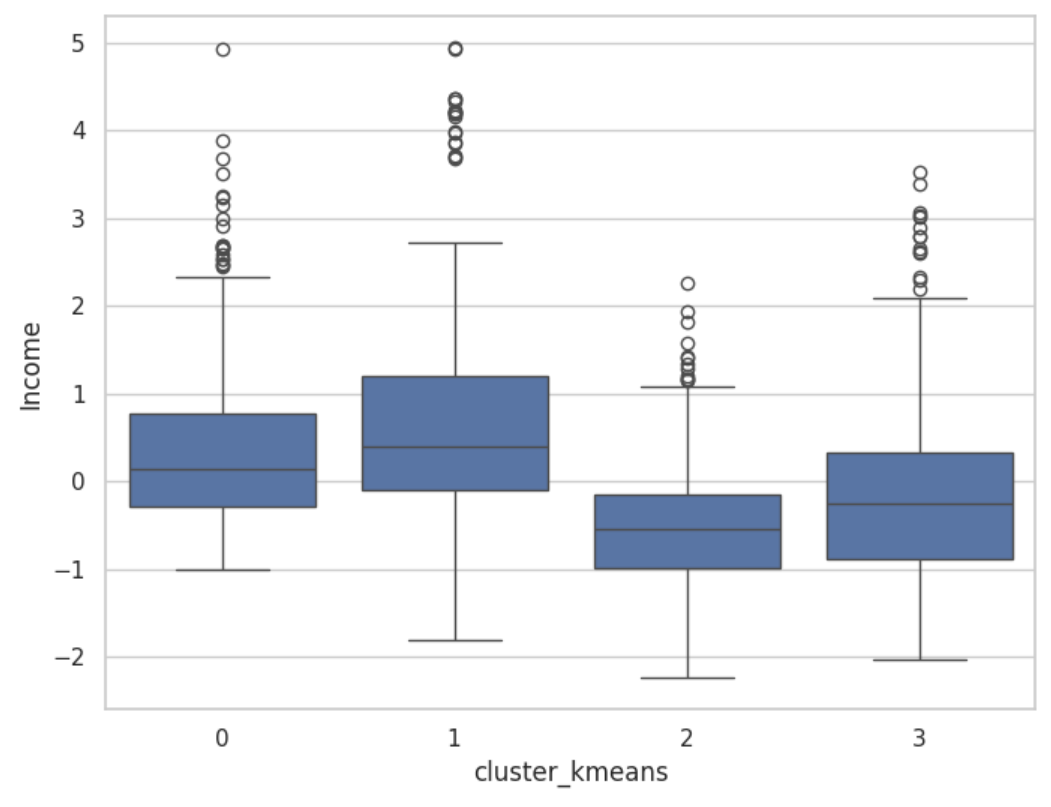
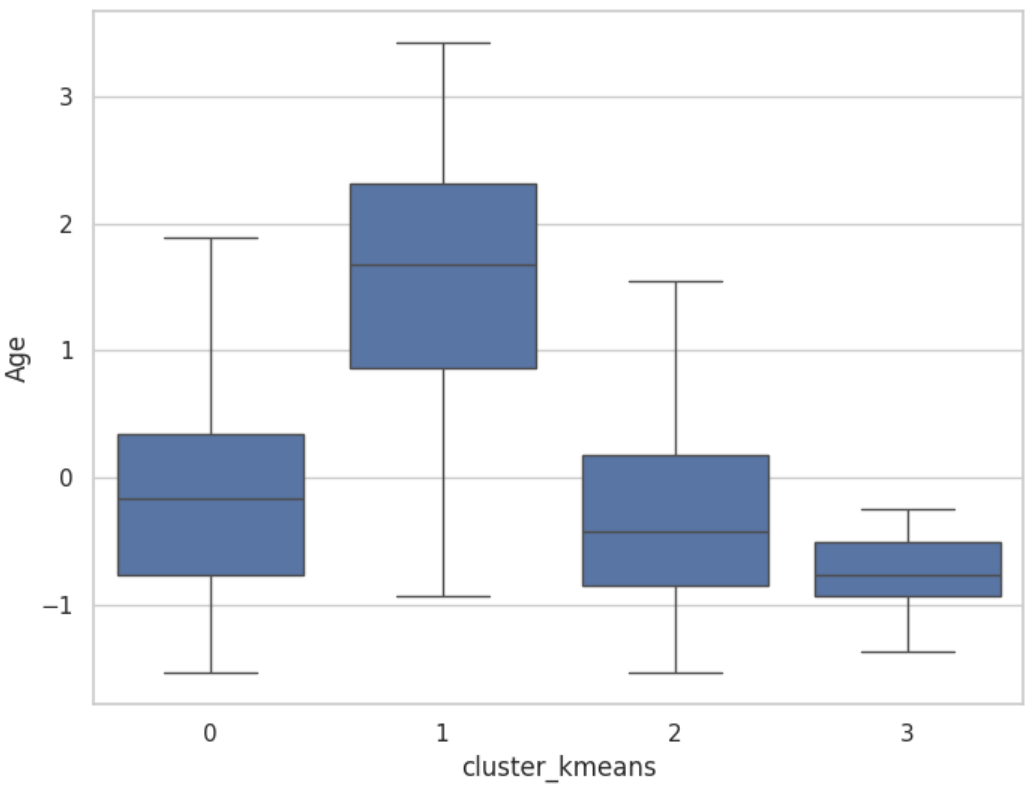
Luego probé con más clústers pero se complejiza mucho y comienza a subdividir clústers que están muy dispersos y son muy pequeños, es posible que el algoritmo no encuentre beneficios significativos al subdividir más.

Por esto, voy a intentar enfocarme más en qué caracteriza a cada clúster:





Para buscar las características principales miré no solo los promedios sino también los promedios de todo el dataset. No podemos sacar muchas conclusiones sabiendo que el promedio de marital status para un clúster es 0,6 sin saber el promedio del dataset completo.



**clúster 0** (rojo)

Ambos géneros pero mayormente hombres de nivel económico medio, secundaria completa y empleados de grandes y medianas ciudades.

**clúster 1** (verde)

Ambos géneros de edad avanzada, casados, nivel económico medio-alto. Atendieron a la universidad. Predominan las mujeres

**clúster 2** (azul)

Adultos jóvenes mayormente mujeres, secundaria completa pero desempleadas. De ciudades pequeñas.

**clúster 3** (amarillo)

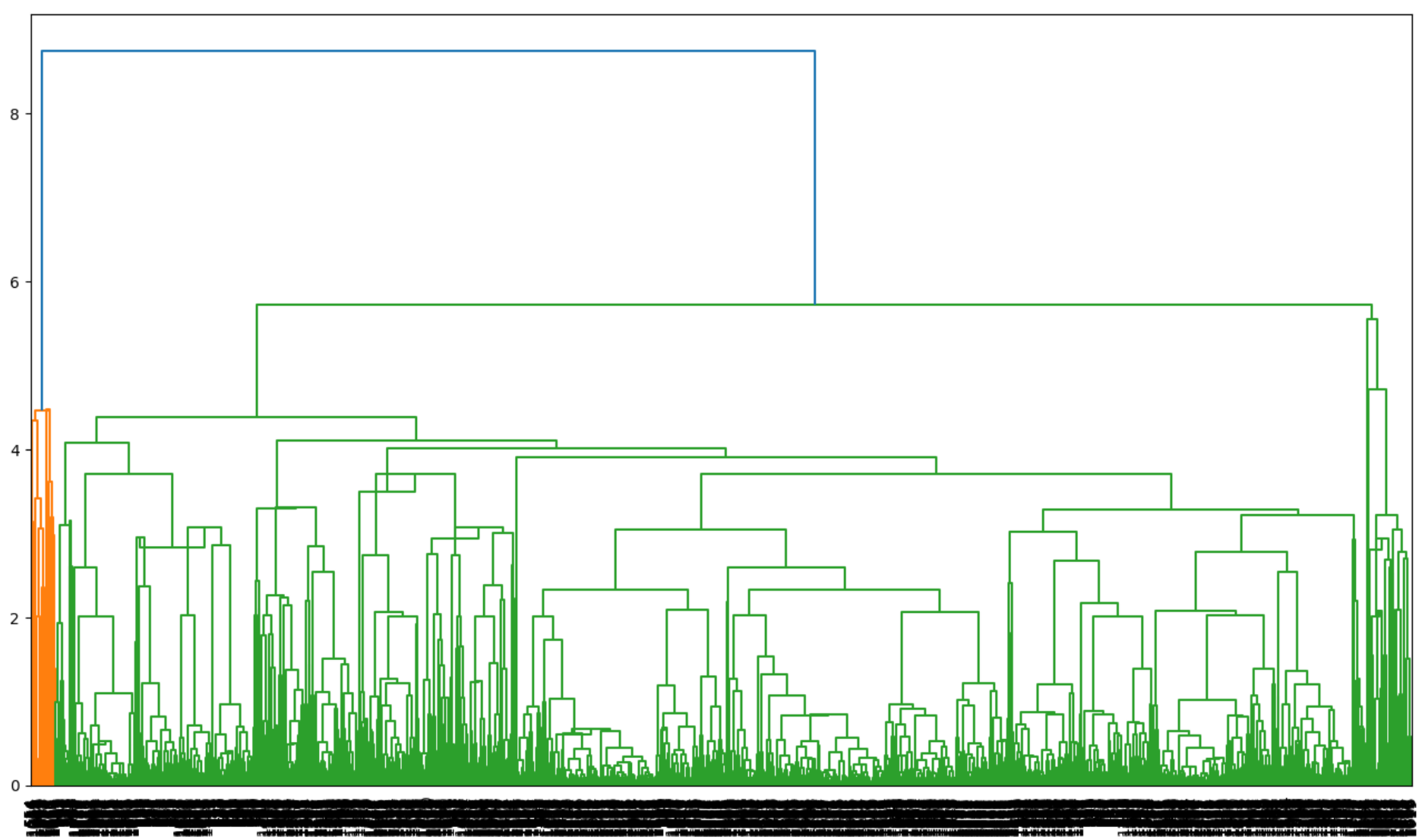
Hombres jóvenes/adolescentes solteros de todos los niveles económicos.

#### 

#### **Aplicación de clústering Jerárquico**

El principal problema de K-means es que tenemos que escoger el número óptimo de clústers K, y en casos reales este no es un problema fácil de resolver. Una alternativa a K-means, es el clústering jerárquico (Hierarchical clústering). En el clúster jerárquico, no hace falta escoger el número de clústers.

Se forman tantos grupos como individuos existan Luego, se agrupan los individuos en base a una medida de similitud entre ellos. Se van agrupan los clústers en base a la distancia que exista entre ellos formando el árbol Dendograma hasta que sólo queda un clúster.



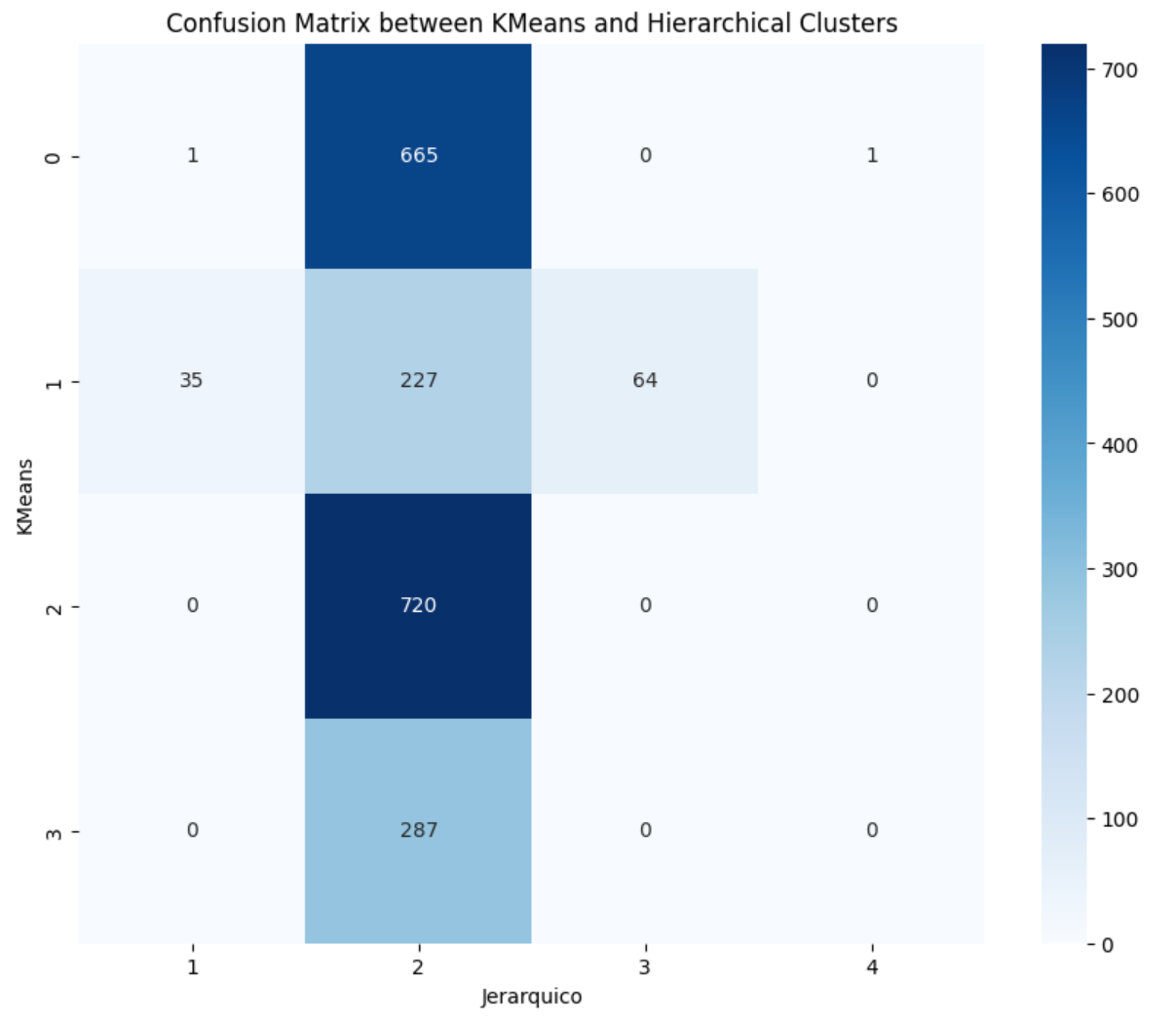
El eje Y contiene la medida de similitud entre observaciones. Mientras más arriba en el eje Y, más diferentes son dos observaciones.

La altura a la que cortemos el dendrograma nos da el número de clústers. Los clústers se forman de manera que las distancias entre las observaciones de cada uno de ellos son menores que el umbral.

Vemos 3 posibles cortes: 4.2, 5 y 6.2 (5, 4 y 3 clústers).

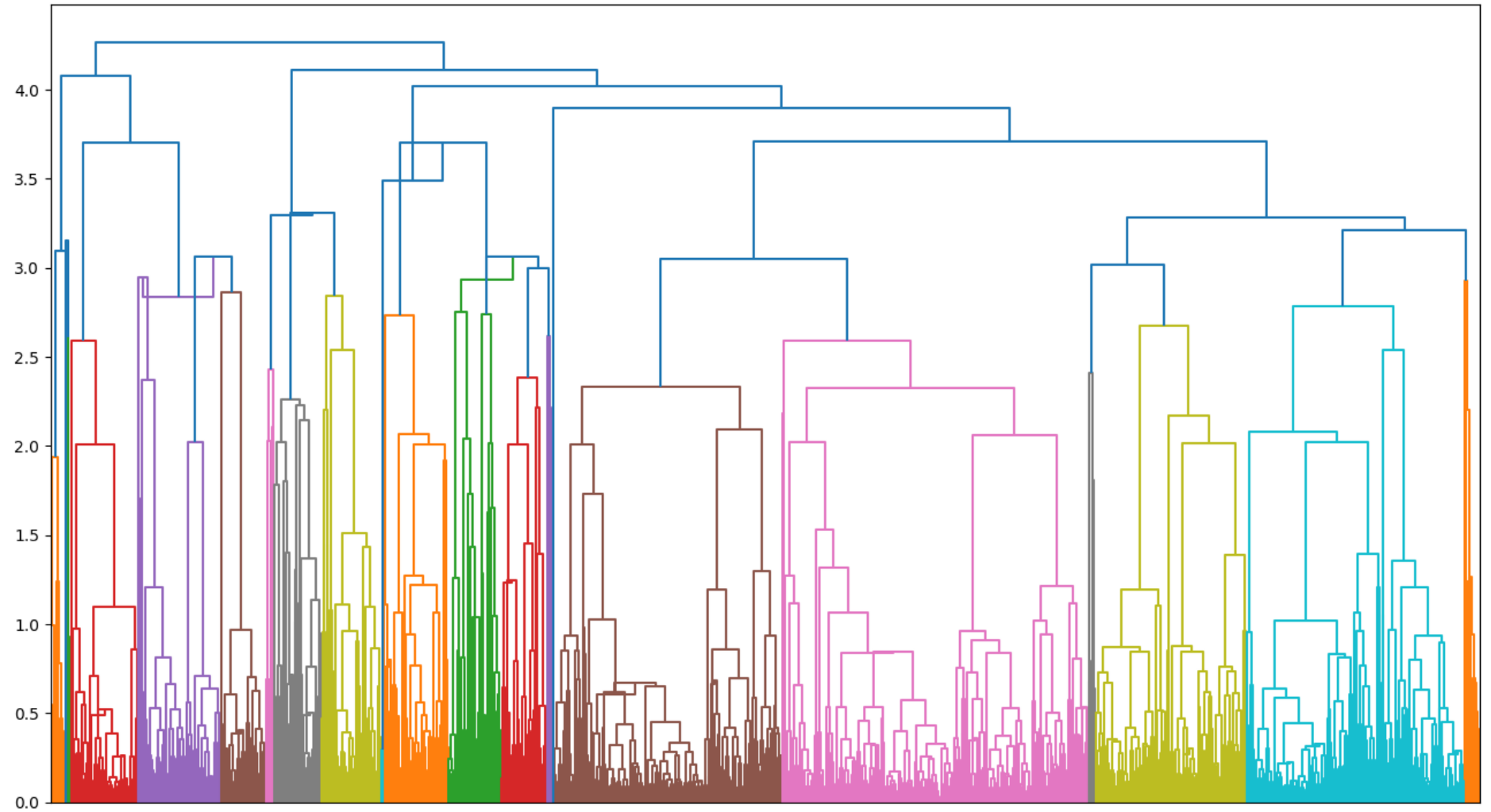
#### 

El problema que estamos viendo acá es que hay un grupo que engloba casi 1900 de los 2000 registros que tenemos. Para poder hacer un buen análisis, vamos a marcarlos como outliers y continuar el análisis con el grupo mayoritario.

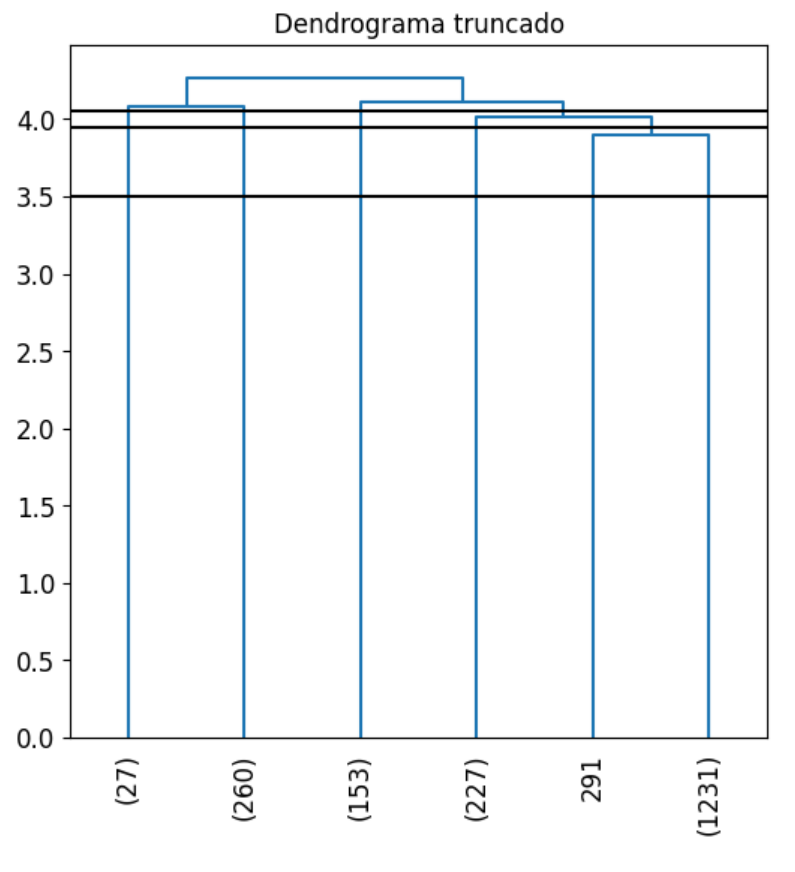


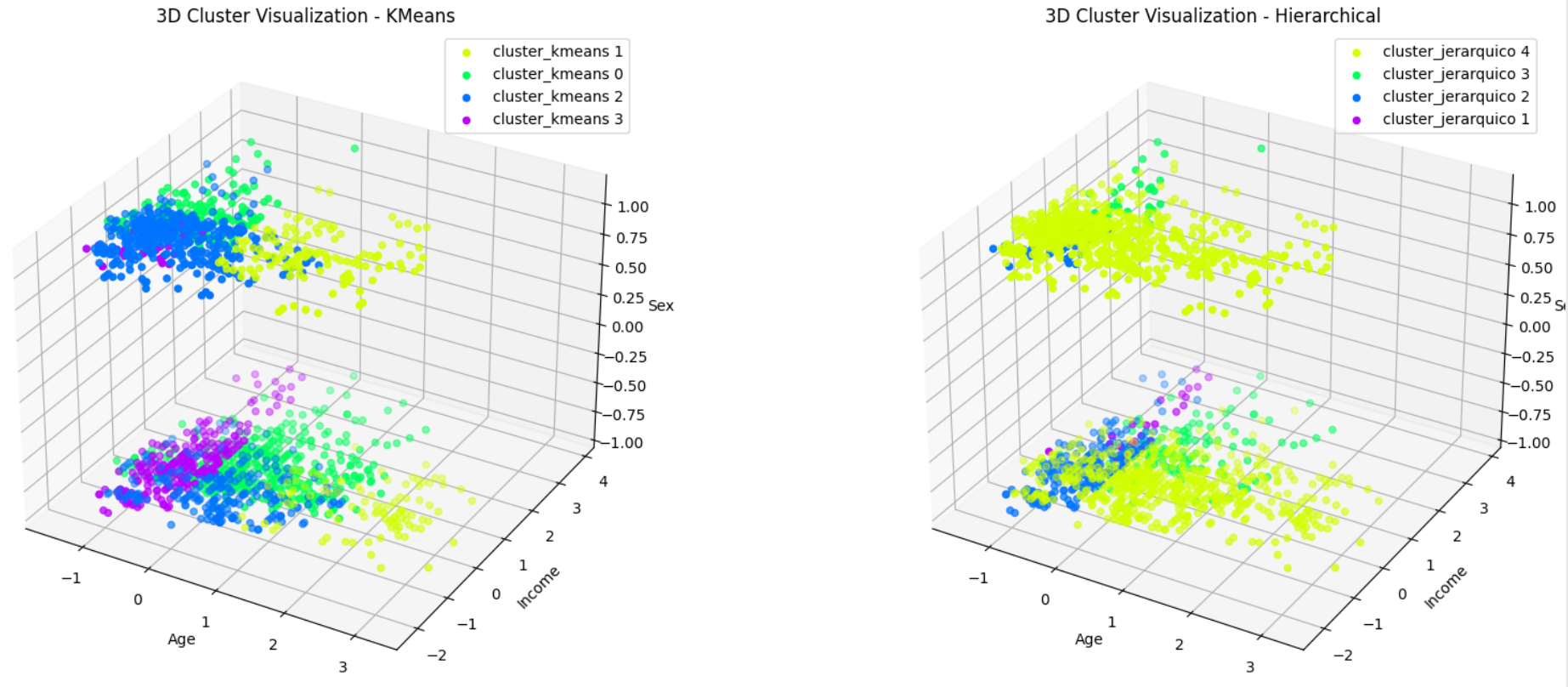
Vemos acá que el 2 es el clúster que agrupa casi todos los clientes y nos vamos a centrar en este.

Vamos a repetir el proceso para poder indagar en clústerizar este gran grupo y asumir el resto como outliers por el momento.

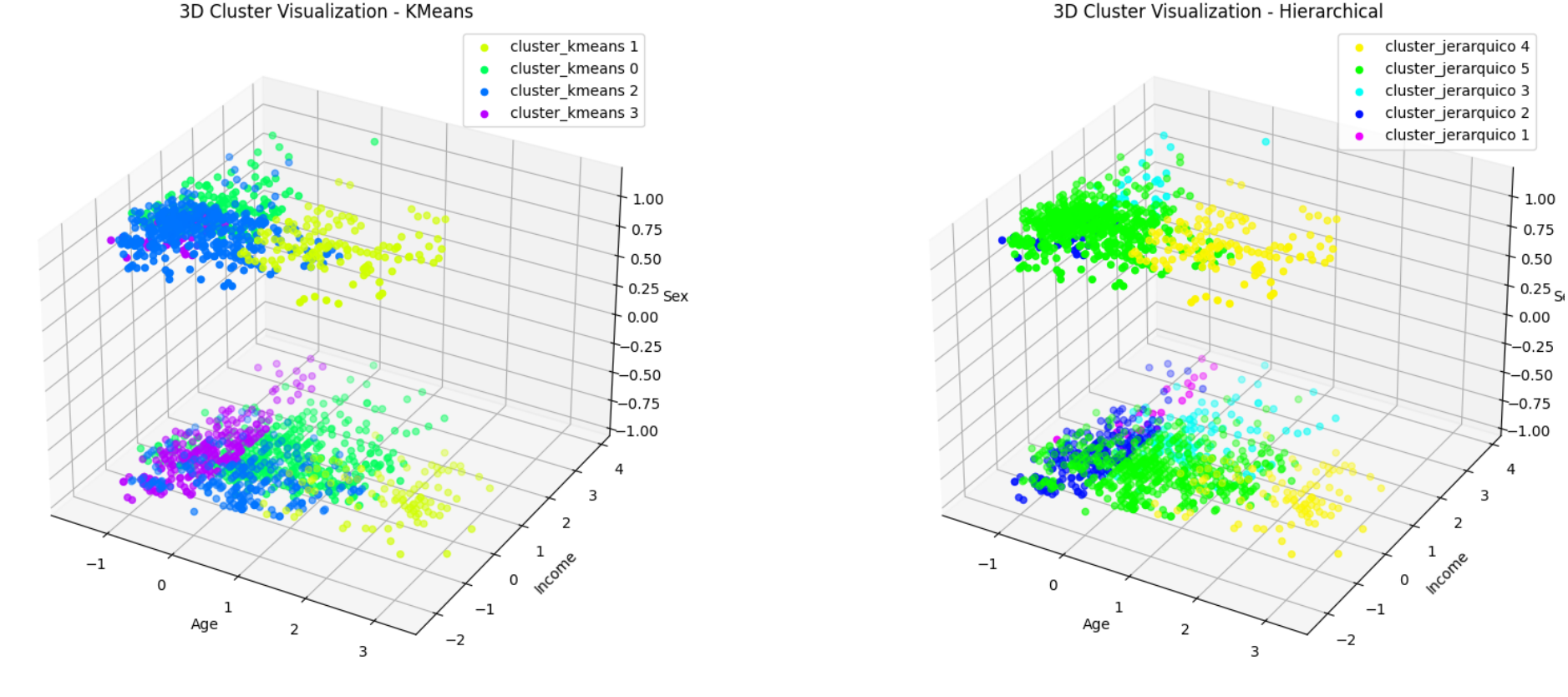


Parecen haber grupos más parejos, no uno que contenga el 95% de los datos. Vamos a cortar cuatro clústers.





Separamos en 4 clústers pero vemos que uno es muy pequeño y apenas tiene 27 registros así que vamos a probar con 5 que divide más.



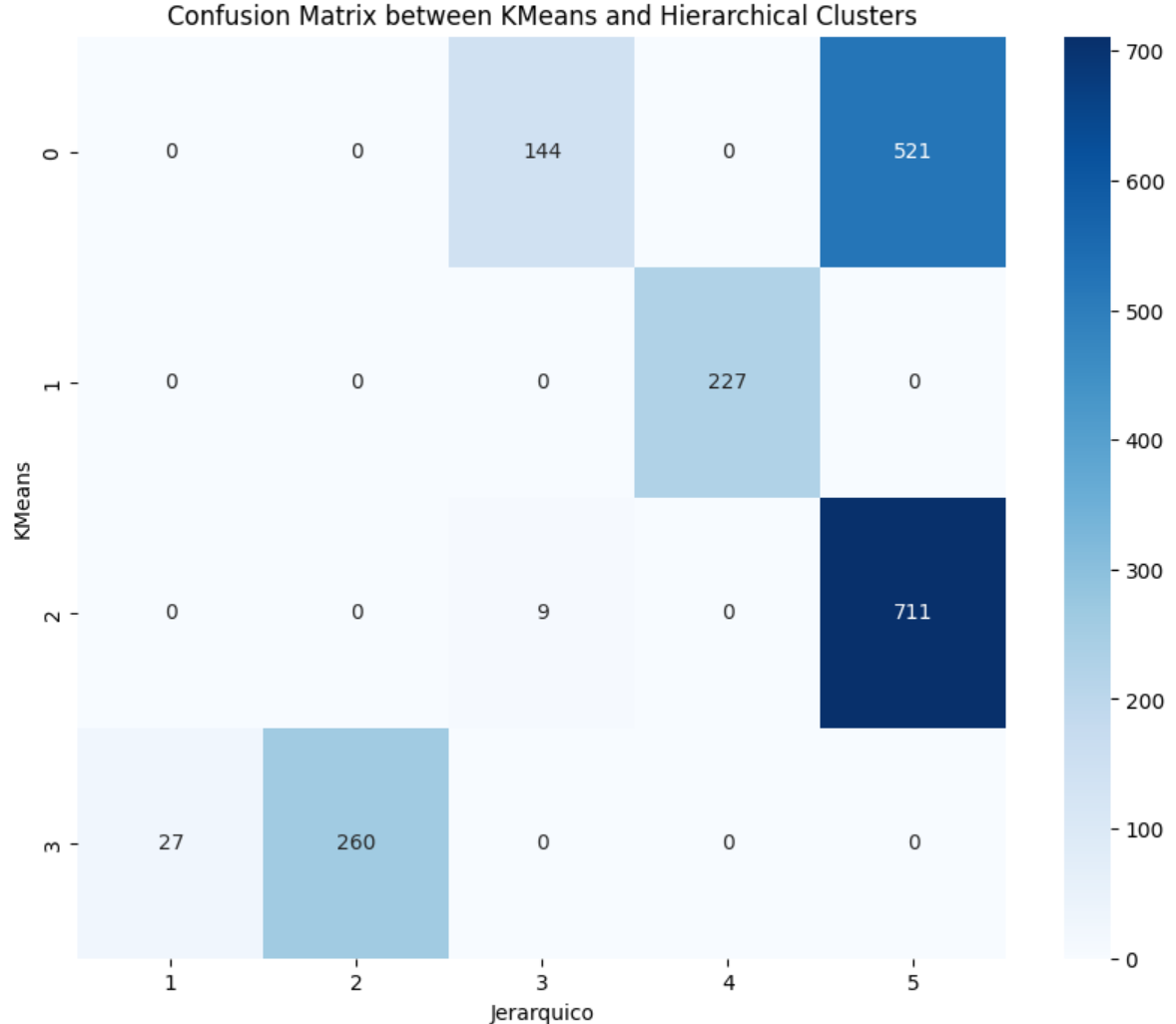
Graficando 5 clústers vemos una gran similitud con el resultado de aplicar Kmeans. El primer clúster apenas tiene registros pero junto con el segundo parecen conformar el tercer clúster de Kmeans.

El tercer clúster es parecido al clúster 0 de la izquierda.

El cuarto clúster es muy similar al primero de Kmeans que abarca a las personas mayores.

El quinto clúster es similar al segundo de Kmeans pero también toma algunos del clúster 0.

Estamos teniendo indicaciones de que se dividen de forma similar para ambos modelos. Podría indicar que es una buena segmentación.



Vemos evidentemente lo que estábamos diciendo sobre el quinto clúster que incluye prácticamente todos los registros del segundo clúster y se lleva también del 0.

Recordemos

* 0: Ambos géneros pero mayormente hombres de nivel económico medio, secundaria completa y empleados de grandes y medianas ciudades.
* 2: Adultos jóvenes mayormente mujeres, secundaria completa pero desempleadas. De ciudades pequeñas.

#### **Aplicación de DBSCAN**

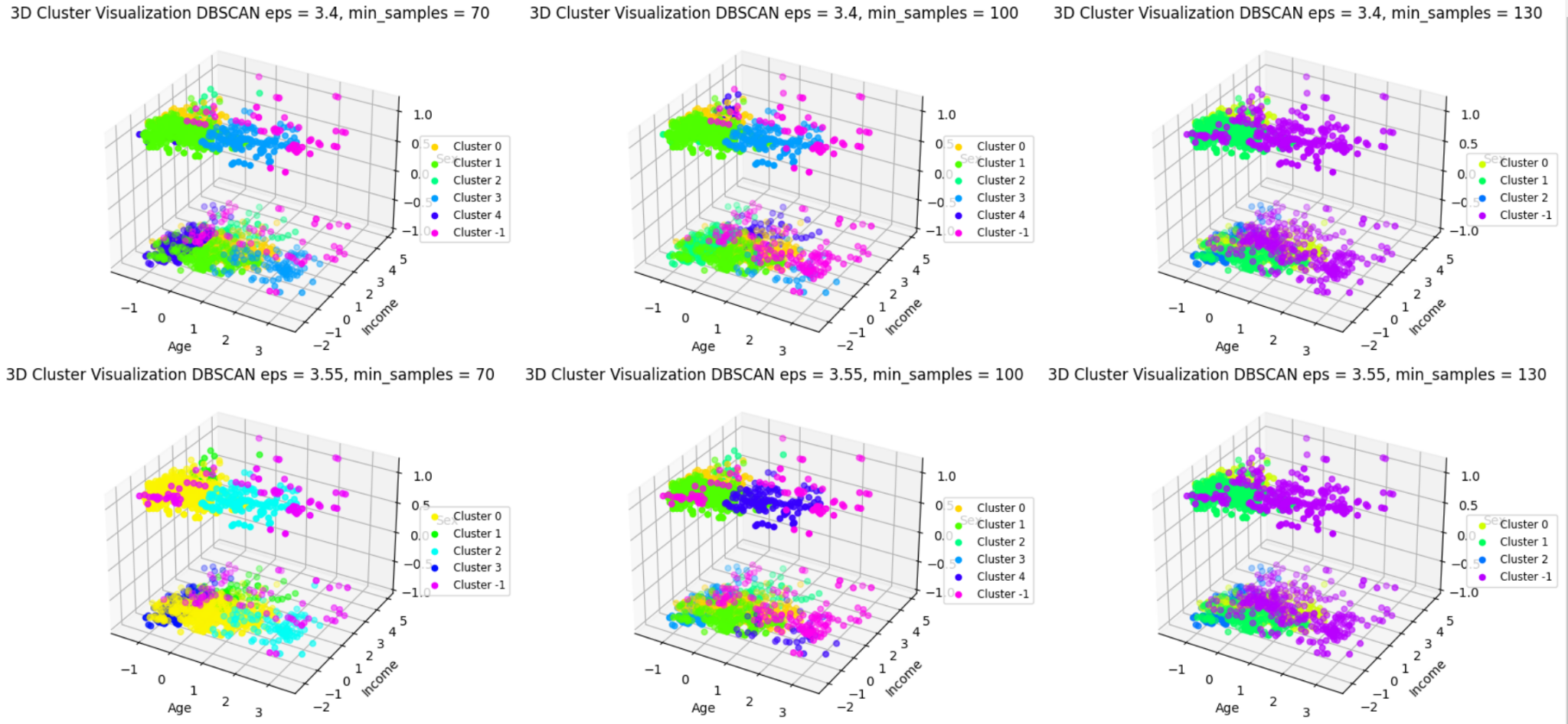
Otro algoritmo de clústering que resuelve algunas de las limitaciones que presenta K-means es DBSCAN.

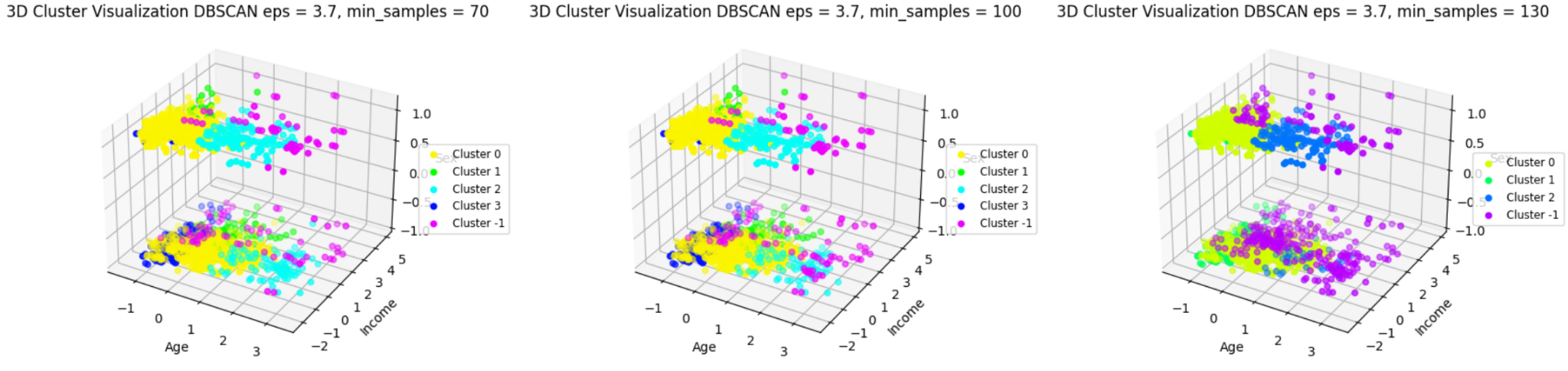
El agrupamiento espacial basado en densidad de aplicaciones con ruido (DBSCAN) considera los clústeres como **áreas de alta densidad** que se miden por el número de muestras próximas a un punto dado. O sea que es un algoritmo que clústeriza por la densidad de puntos en el espacio. Es utilizado para identificar clústers de cualquier forma en un conjunto de datos que contiene ruido y valores atípicos.

La idea clave es que, para cada punto de un clúster, la vecindad de un radio (epsilon) dado tiene que contener al menos un número mínimo de puntos (min\_samples).

Como no sabemos qué epsilon y min samples usar, hacemos un for que recorra varias combinaciones.

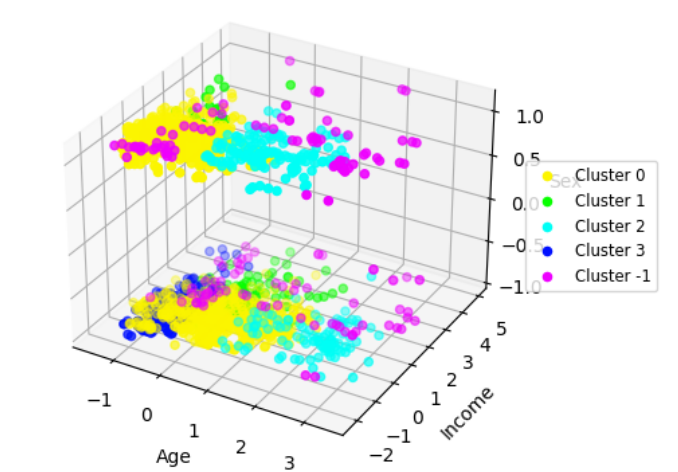
* eps\_values = [3.4, 3.55, 3.70]
* min\_samples\_values = [70, 100, 130]





Algunos de los más interesantes son los siguientes:

Eps: 3.55 Min\_samples: 70



Eps: 3.7 Min\_samples: 70 o 100 (son casi iguales)



Lo interesante de este modelo, y la nueva mirada que nos da es que el cluster que en los modelos anteriores que se formaban de adultos mayores es marcado como outliers. Esto podría ser un dato interesante para no desperdiciar tiempo ni dinero en conocer un grupo tan pequeño y disperso.

### **Conclusión General**

#### **Clústeres Identificados**

1. **Clúster 0 (Hombres jóvenes de nivel económico medio):**
   * Características: Mayormente hombres jóvenes, nivel económico medio, secundaria completa, empleados de grandes y medianas ciudades.
   * Estrategias:
     + Ofrecer productos tecnológicos, ropa deportiva, y entretenimiento.
     + Promociones en productos de conveniencia y ahorro de tiempo.
     + Campañas de marketing en redes sociales y aplicaciones móviles.
2. **Clúster 1 (Adultos mayores, mayoritariamente mujeres, nivel económico medio-alto):**
   * Características: Ambos géneros pero predominan las mujeres, edad avanzada, casados, nivel económico medio-alto, universitarios.
   * Estrategias:
     + Ofrecer productos de salud y bienestar, alimentos saludables y orgánicos.
     + Programas de fidelización y descuentos para jubilados.
     + Marketing a través de medios tradicionales como la televisión y correo directo.
3. **Clúster 2 (Mujeres jóvenes de clase media-baja):**
   * Características: Adultos jóvenes mayormente mujeres, secundaria completa, desempleadas, de ciudades pequeñas.
   * Estrategias:
     + Promociones y descuentos en productos básicos y de primera necesidad.
     + Ofrecer programas de capacitación y empleo.
     + Marketing a través de redes sociales y promociones locales.
4. **Clúster 3 (Hombres jóvenes/adolescentes solteros de todos los niveles económicos):**
   * Características: Hombres jóvenes/adolescentes solteros, diversos niveles económicos.
   * Estrategias:
     + Ofrecer productos de moda, tecnología y entretenimiento.
     + Promociones en productos de conveniencia y ahorro de tiempo.
     + Marketing a través de influencers y redes sociales.

#### **Consideración de Outliers**

El análisis con DBSCAN identificó un grupo significativo de outliers, principalmente adultos mayores, que no se agrupan claramente en ningún clúster significativo.

* **Estrategia para Outliers:**
  + Evaluar el costo-beneficio de atender a este grupo disperso.
  + Considerar estrategias generales que puedan atraer a una audiencia amplia y diversa sin enfocarse demasiado en los outliers.
  + Podría ser más eficiente enfocar los esfuerzos de marketing y desarrollo de productos en los clústeres más grandes y definidos.

La segmentación de clientes ha revelado patrones claros que pueden guiar nuestras estrategias de marketing y desarrollo de productos. Al enfocar nuestros recursos en los clústeres más grandes y definidos, podemos ofrecer experiencias de compra más personalizadas, mejorar la satisfacción del cliente y, en última instancia, aumentar la rentabilidad. Los outliers, aunque importantes, podrían no justificar una inversión significativa dada su dispersión y tamaño reducido. Por lo tanto, nuestra estrategia debe centrarse en los segmentos clave identificados para maximizar el retorno de la inversión.