Universidad de La Laguna

Memoria Práctica Agrupamiento

Tratamiento Inteligente de Datos (TID)

Raúl González Acosta

Grado en Ingeniería Informática
Dpto. de Ingeniería Informática y de Sistemas
Escuela Superior de Ingeniería y Tecnología
Universidad de La Laguna

La Laguna, a 22 de marzo de 2025

Índice general

1

Descripción de la Base de Datos

1.1. Descripción de la base de datos

La base de datos que usaremos para este análisis proviene de un centro comercial que busca entender mejor a sus clientes para mejorar las estrategias de ventas y la personalización de los servicios que ofrecen. Este conjunto de datos recopila información de 200 clientes, proporcionando una muestra significativa para realizar el análisis de segmentación. En el cuadro 1.1 se describen las 5 variables del fichero *mallCustomers.csv*

Variable	Descripción	Valores
CustomerID	Identificador del cliente	Numérico
Gender	Género del cliente	Male/Female
Age	Edad del cliente	Numérico
AnnualIncome	Ingresos anuales del cliente	Numérico
SpendingScore	Puntuación dada al cliente por el supermercado	Numérico

Cuadro 1.1: *Descripción de las variables a analizar de mallCustomer.csv*

Si visualizamos un espectro de los datos haciendo uso de un histograma, podemos ver que los clientes del centro comercial tienen edades comprendidas entre los 18 y 70 años. En cuanto a los términos económicos, presentan un ingreso anual medio de aproximadamente 61 mil unidades monetarias. Además, poseen una puntuación media de gasto de 50 puntos. Estas observaciones ofrecen una visión general del perfil promedio de los clientes. Sin embargo, en este informe buscamos realizar una tarea de agrupamiento que nos permita ir más allá de este perfil promedio, identificando grupos específicos de clientes con características y comportamientos similares.

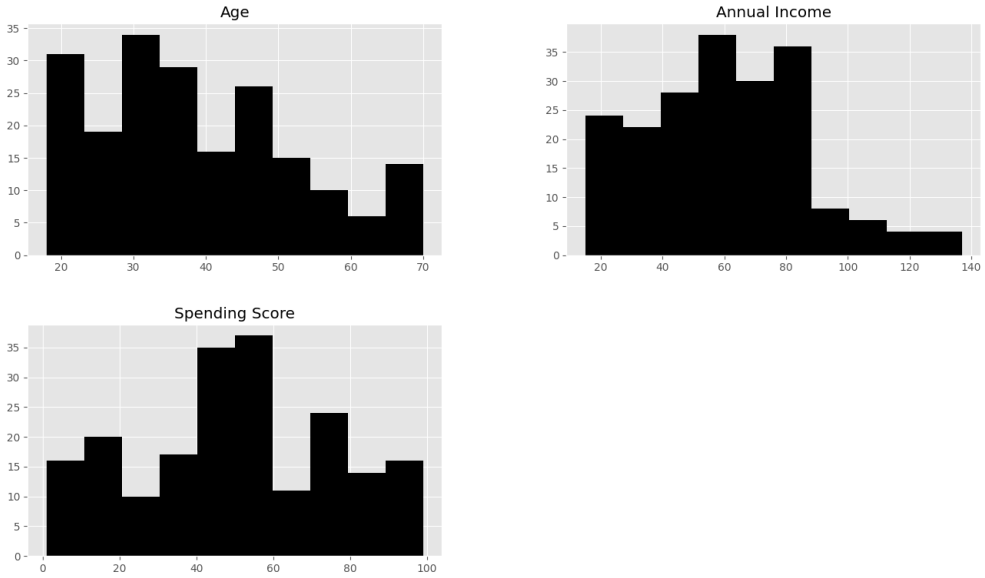

Figura 1.1: *Histograma de las variables numéricas de la base de datos mallCustomers.csv*

1.2. Tratamiento de los datos

Antes de proceder con las tareas de agrupamiento, debemos preparar adecuadamente los datos. Afortunadamente, la base de datos es completa, por tanto no presenta valores ausentes, en consecuencia no tenemos necesidad de realizar imputaciones o eliminaciones de datos. Sin embargo, para garantizar que los algoritmos de agrupamiento funcionen de manera óptima, tenemos que estandarizar o normalizar los datos. En este caso, se ha optado por la estandarización[1], puesto que con esta estrategia se trata de evitar interpretaciones erróneas de los ingresos como "0", lo cual es poco representativo de la realidad de nuestros datos. Además, para los agrupamientos, eliminaremos la variable "*CustomerID*"puesto que no aporta ninguna información relevante a la hora de realizar los agrupamientos.
En cuanto a los valores atípicos o *outliers* que se observan en la figura 1.2, debemos evaluar el impacto que estos tienen en los resultados de los agrupamientos. Tras realizar una comparativa, observamos que la calidad del agrupamiento mejora al conservarlos [2] para el análisis. Esto sugiere que, en este contexto particular, los *outliers* pueden contener información valiosa para la segmentación de clientes.

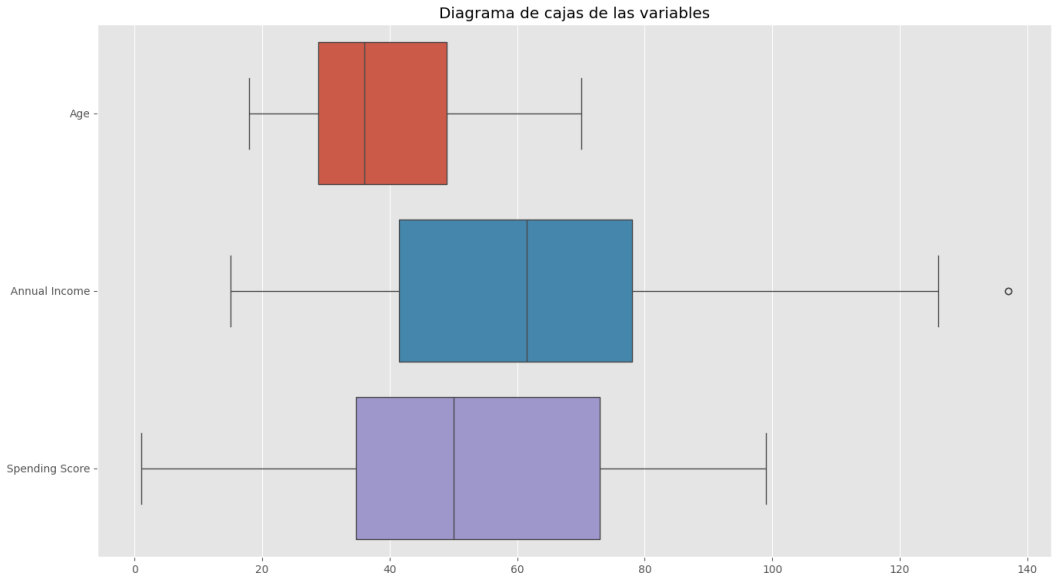

Figura 1.2: *Diagrama de cajas de las variables numéricas de la base de datos mallCustomers.csv*

[1] El código para normalizar también se encuentra disponible en el cuaderno Jupyter
[2] El código para excluirlos también se encuentra disponible en el cuaderno Jupyter

2

Modelos de agrupamiento

Para construir nuestros modelos de agrupamiento, haremos dos aproximaciones, en la primera no segregaremos por el sexo de los clientes, mientras que en la segunda trataremos de hacer otra aproximación de forma que los separaremos para un mejor estudio del comportamiento de los clientes.

2.1. Agrupamiento *K-Means*

Para determinar el número óptimo de clusters para nuestro agrupamiento, implementamos el método del codo. Si nos fijamos en la figura 2.1, donde la gráfica realiza el "codo", esta nos sugiere que el número adecuado de clusters se encuentre entre 5 y 7. Tras experimentar con estos tres valores enteros, decidimos optar por 6 clusters, puesto que esta configuración nos proporciona los resultados más coherentes y significativos.

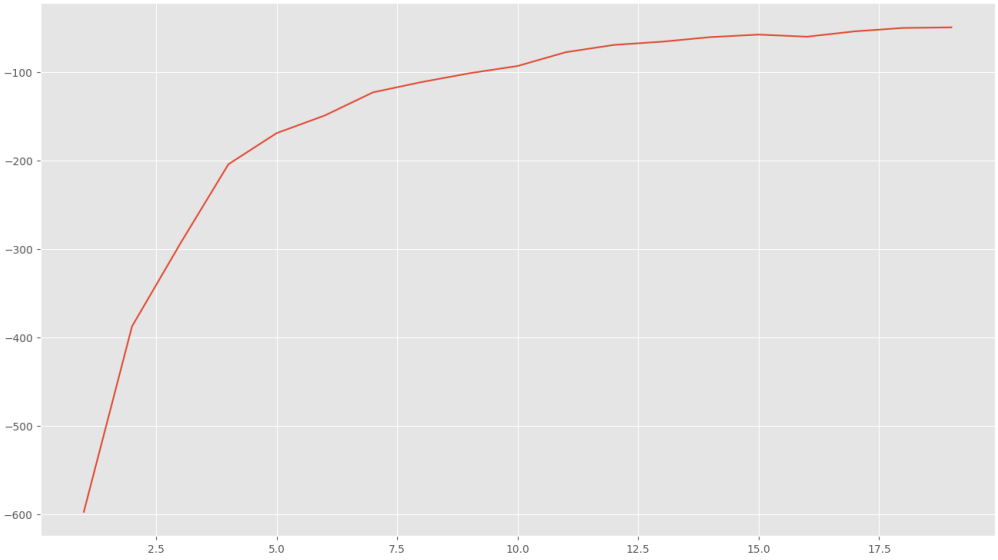

Figura 2.1: *Gráfica de los K-Means de los clusters*

Realizando *K-Means* con los 6 clusters previamente comentados, podemos extraer las siguientes conclusiones de los grupos generados:

- **Equilibristas financieros**: Con una edad promedio de 27.51 años, ingresos anuales de 50.10 mil y una puntuación de gasto de 43.62, estos consumidores representan un equilibrio entre ingresos y gastos, posiblemente jóvenes profesionales con estabilidad financiera que priorizan el consumo moderado y el ahorro.
- **Epicúreos modernos**: Este grupo tiene una edad promedio de 32.85 años, ingresos anuales altos (80.03 mil) y una puntuación de gasto elevada (82.18). Se trata de clientes con un buen poder adquisitivo, estando en el punmto más álgido de su carrera, que tienen una fuerte inclinación al consumo, haciendo un estilo de vida activo y con gastos en experiencias y bienes de lujo.
- **Ultraconservadores selectivos**: Con una edad promedio de 41.94 años, ingresos elevados (88.94 mil) y una puntuación de gasto baja (16.97), estos clientes podrían representar a profesionales con alto poder adquisitivo que son cautelosos con sus gastos, posiblemente más orientados a inversiones y ahorro.
- **Ahorradores prudentes**: Este grupo tiene una edad promedio de 55.55 años, ingresos anuales medios (48.48 mil) y una puntuación de gasto de 41.78. Son clientes que probablemente tienen estabilidad financiera, pero priorizan el ahorro o gastos específicos y moderados.
- **Jóvenes oportunistas**: Con una edad promedio de 25.27 años, ingresos anuales bajos (25.73 mil) y una puntuación de gasto muy alta (79.36), este grupo podría estar compuesto por jóvenes que priorizan el consumo a pesar de tener ingresos limitados, posiblemente estudiantes o jóvenes profesionales en etapas iniciales de sus carreras.
- **Élite del gasto**: Este grupo tiene una edad promedio de 33 años, ingresos anuales elevados (114.71 mil) y una alta puntuación de gasto (78.43). Representan a clientes con alto poder adquisitivo y una fuerte tendencia a gastar, posiblemente en productos de lujo, moda y entretenimiento.

A continuación se muestra en la figura 2.2, la representación espacial de los *clusters* generados por *K-Means*.

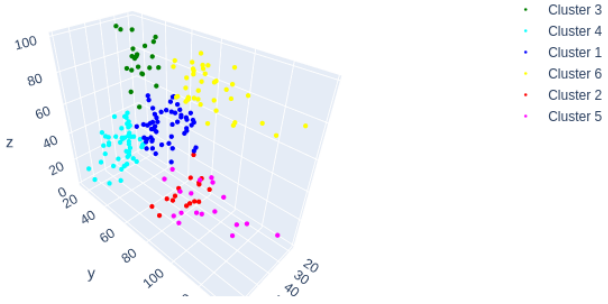

Figura 2.2: *Clusters generados por el agrupamiento de K-Means*

2.2. Dendogramas

Ya hemos visto previamente, un modelo de agrupamiento que nos ha permitido ver los distintos grupos en base a los comportamientos que tienen los clientes. Sin embargo, en este apartado queremos poder desgranar un poco más esos grupos, de forma que haremos uso de tres tipos distintos de dendogramas: método *single*, método *centroid*, y método *ward*.

2.2.1. Método *Single*

A continuación en la figura 2.3 se observa una estructura del diagrama alargada, lo que indica que el método de enlace simple (*single*) tiende a generar clústeres con relaciones más cercanas entre ciertos clientes, pero menos compactos. Además, la presencia de cadenas largas sugiere que los grupos formados pueden ser más difusos y no tan homogéneos, por tanto puede haber agrupaciones menos representativas y con outliers. Por último se ve, como en la base hay diferenciados los 6 grupos que había mencionado ya previamente.

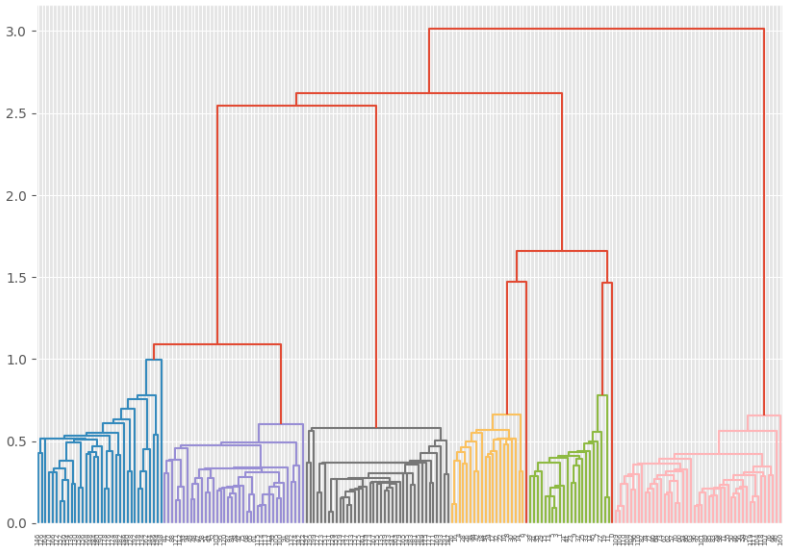

Figura 2.3: *Dendograma haciendo uso del método single*

2.2.2. Método *Centroid*

En la figura 2.4 se observa como este método agrupa los clientes según la distancia de sus centroides, lo que puede generar clústeres más dispersos y menos balanceados. Se nota que hay algunas fusiones abruptas, lo que indica que ciertos grupos tienen una alta variabilidad interna, por ende es posible que los clústeres representen patrones de consumo más extremos. En definitiva en este dendograma tambíen podemos ver, como en la base hay diferenciados los 6 grupos de manera clara.

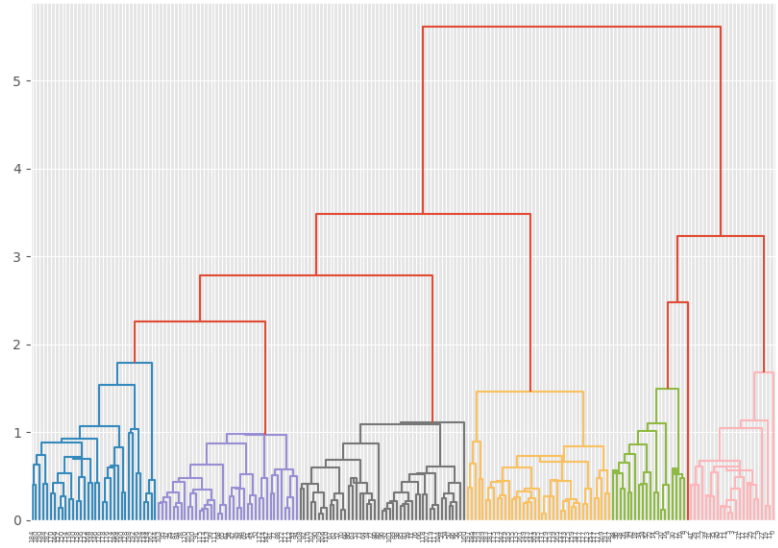

Figura 2.4: *Dendograma haciendo uso del método centroid*

2.2.3. Método *Ward*

En la figura 2.5 permite observar como este método minimiza la varianza dentro de los clústeres, lo que produce grupos más homogéneos y equilibrados. Se observan fusiones progresivas y diferenciadas, lo que indica que los clientes se agrupan en conjuntos bien definidos. Además en este dendograma podemos observar que se diferencia más bien en 7 grupos distintos, lo que sugiere patrones de consumo bien diferenciados.

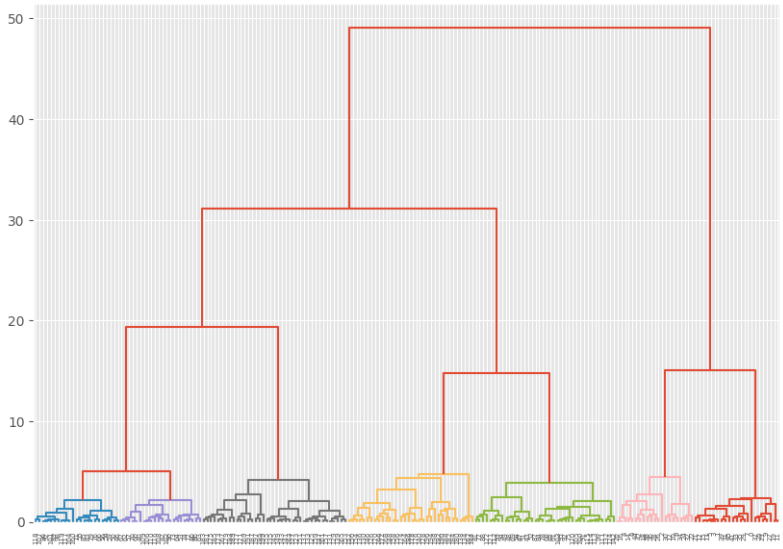

Figura 2.5: *Dendograma haciendo uso del método ward*

2.2.4. Conclusiones

A continuación en el cuadro 2.1 podemos observar una tabla con los resultados del índice de *cophenet* obtenidos con los distintos dendogramas:

Método	Índice de *cophenet*
Single	0.81
Centroid	0.94
Ward	0.95

Cuadro 2.1: *Resultados del índice de cophenet de los dendogramas*

Con esto podemos concluir que tanto el método *centroid* como el *ward* ofrecen una alta fidelidad en la representación de las distancias originales entre los puntos de datos, con índices de 0.94 y 0.95, respectivamente. Aunque el método *ward* muestra ligeramente la mejor preservación de las distancias, el método *centroid*, con un enfoque más equilibrado y un similar índice de correlación, proporciona una estructura de agrupamiento que se considera más apropiada para nuestros objetivos analíticos. Es por ello, que se ha decidido optar por la segmentación de clientes proporcionada por el método *centroid*. Bajo ningún circunstancia, el método *single* es recomendable, ya que si se busca una estructura de agrupamiento confiable, su bajo índice de correlación (0.81) sugiere una mala representación de la distancia real entre puntos.

2.3. Agrupamiento por género

Para esta segunda aproximación, segregaremos nuestra base de datos en dos distintas: uno para el género masculino y otro para el género femenino. Los datos se han tratado de la misma forma que en el caso previo, es decir, tratandolos bajo una estandarización y manteniendo los outliers, puesto que estos lograban mejores resultados.

2.3.1. Agrupamiento exclusivo del género femenino

Realizando *K-Means* con los mismos 6 clusters que en el caso general, podemos extraer las siguientes conclusiones de los grupos generados:

- **Jóvenes con bajo poder adquisitivo y alto gasto**: Este grupo tiene una edad promedio de 25.46 años, ingresos anuales bajos (25.69 mil) y una puntuación de gasto muy alta (80.54). Son mujeres jóvenes con ingresos limitados, pero con una fuerte tendencia a gastar, lo que sugiere un enfoque hacia el consumo inmediato, posiblemente en entretenimiento, moda o productos de cuidado personal.
- **Profesionales de altos ingresos pero bajo gasto**: Con una edad promedio de 43.79 años, ingresos elevados (93.29 mil) y una puntuación de gasto baja (20.64), este grupo representa a mujeres profesionales con alto poder adquisitivo que son muy cautelosas con sus gastos. Probablemente priorizan el ahorro, la inversión o el bienestar a largo plazo sobre el consumo de bienes o servicios.
- **Clientes de alto poder adquisitivo y alto gasto**: Este grupo tiene una edad promedio de 32.19 años, ingresos anuales altos (86.05 mil) y una puntuación de gasto elevada (81.67). Se trata de mujeres con altos ingresos y una fuerte inclinación al consumo, probablemente mujeres jóvenes y exitosas que disfrutan de un estilo de vida activo, con compras de lujo y experiencias exclusivas.
- **Adultos mayores con ingresos y gasto moderado**: Con una edad promedio de 54.15 años, ingresos anuales medios (54.23 mil) y una puntuación de gasto de 48.96, este grupo probablemente está formado por mujeres con estabilidad financiera que priorizan el consumo moderado. Mantienen un enfoque equilibrado entre el gasto en necesidades básicas y el ahorro para el futuro.
- **Adultos con ingresos bajos y bajo gasto**: Este grupo tiene una edad promedio de 41.54 años, ingresos anuales bajos (26.54 mil) y una puntuación de gasto baja (20.69). Son mujeres que probablemente adoptan un enfoque conservador tanto en sus ingresos como en sus gastos, con un enfoque en el ahorro y una gestión cuidadosa de sus recursos financieros.
- **Jóvenes con ingresos moderados y gasto moderado**: Con una edad promedio de 27.96 años, ingresos anuales de 57.36 mil y una puntuación de gasto de 47.12, este grupo representa a mujeres jóvenes profesionales que mantienen un equilibrio entre ingresos y gastos. Son consumidoras que buscan estabilidad financiera mientras disfrutan de un consumo moderado y responsable.

Usando el método *centroid*, en la figura 2.6 se ven los clusters generados resaltando en negro los centroides de dichas agrupaciones.

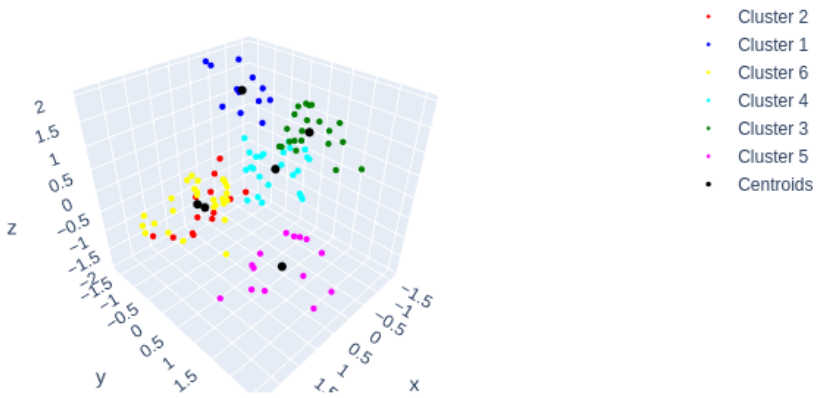

Figura 2.6: *Clústeres de la segregación femenina*

2.3.2. Agrupamiento exclusivo del género masculino

Realizando *K-Means* con los mismos 6 clusters que en el caso general, podemos extraer las siguientes conclusiones de los grupos generados:

- **Jóvenes de alto poder adquisitivo y alto gasto**: Este grupo tiene una edad promedio de 33.56 años, ingresos anuales elevados (81.56 mil) y una puntuación de gasto muy alta (83.19). Son hombres con un buen nivel adquisitivo y una fuerte inclinación al consumo, probablemente profesionales exitosos que gastan en experiencias, tecnología, moda y entretenimiento.
- **Adultos mayores con ingresos y gasto moderado**: Con una edad promedio de 58.84 años, ingresos anuales medios (47.80 mil) y una puntuación de gasto de 41.00, este grupo representa a hombres que, aunque cuentan con estabilidad financiera, son más prudentes con sus gastos, priorizando necesidades esenciales y ahorro.
- **Profesionales de altos ingresos pero bajo gasto**: Este grupo tiene una edad promedio de 32.50 años, ingresos muy altos (119.50 mil) y una puntuación de gasto baja (41.00). Son hombres con un gran poder adquisitivo, pero que mantienen un control riguroso sobre sus gastos, posiblemente priorizando inversiones o ahorros en lugar del consumo inmediato.
- **Adultos con ingresos altos pero muy bajo gasto**: Con una edad promedio de 40.28 años, ingresos elevados (80.72 mil) y una puntuación de gasto muy baja

(14.17), estos hombres probablemente tienen un enfoque financiero conservador, enfocándose en la acumulación de riqueza y evitando gastos innecesarios.

- **Jóvenes con ingresos bajos y gasto elevado**: Este grupo tiene una edad promedio de 26.85 años, ingresos anuales bajos (25.62 mil) y una puntuación de gasto relativamente alta (65.08). Son hombres jóvenes que, a pesar de contar con bajos ingresos, mantienen un alto nivel de consumo, posiblemente priorizando entretenimiento, ocio y tecnología.
- **Jóvenes con ingresos moderados y gasto equilibrado**: Con una edad promedio de 24.50 años, ingresos anuales de 56.42 mil y una puntuación de gasto de 52.42, estos hombres representan un perfil con una gestión más equilibrada de sus recursos, combinando el consumo con el ahorro para mantener estabilidad financiera.

Usando el método *centroid*, en la figura 2.7 se ven los clusters generados resaltando en negro los centroides de dichas agrupaciones.

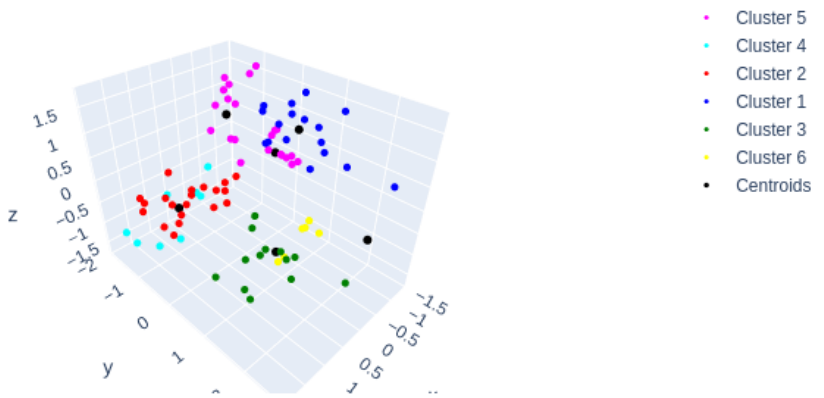

Figura 2.7: *Clústeres de la segregación masculina*

2.3.3. Diferencias entre los grupos segregados por género

Los agrupamientos realizados exclusivos por género muestran patrones distintos que difieren del análisis de agrupamiento original que incluía a todos los clientes sin distinción de este. A continuación, se destacan las principales diferencias entre estos agrupamientos y el agrupamiento realizado originalmente:

- **Niveles de Ingreso y Gasto**
  - **Hombres**: Hay una mayor variabilidad en los ingresos, con un grupo de muy altos ingresos (hasta 119.5 mil) pero con gasto bajo (41.0 puntos). Además, los hombres parecen estar más polarizados entre gastar mucho o gastar muy poco.
  - **Mujeres**: Aunque también hay un grupo de alto poder adquisitivo con alto gasto, la distribución de los ingresos es más equilibrada. Los grupos femeninos tienden a tener una correlación más clara entre ingresos y consumo.
- **Patrón de Gasto**
  - **Hombres**: Existen grupos que, a pesar de tener altos ingresos, gastan muy poco (por ejemplo, el grupo de 40.28 años con ingresos de 80.72 mil pero un gasto de solo 14.17). También hay jóvenes con ingresos bajos que gastan de manera significativa.
  - **Mujeres**: Aunque también hay diferencias en los patrones de gasto, los grupos femeninos muestran una tendencia más balanceada, con menos extremos entre gasto y ahorro.
- **Edad Promedio y Etapas de Vida**
  - **Hombres**: Hay un grupo de adultos mayores (58.84 años) con ingresos y gastos moderados, lo que sugiere estabilidad financiera y prudencia en el consumo.
  - **Mujeres**: El grupo de mayor edad tiene 54.15 años y también mantiene un equilibrio en ingresos y gastos, aunque con una ligera tendencia al consumo más alto en comparación con los hombres de la misma edad.
- **Diferencias en la Prioridad del Consumo**
  - **Hombres**: La existencia de grupos de alto ingreso pero con bajo gasto sugiere que algunos hombres prefieren estrategias de inversión y ahorro en lugar de consumo inmediato. También hay grupos jóvenes con ingresos bajos que gastan mucho, lo que puede sugerir un comportamiento impulsivo o enfocado en experiencias a corto plazo.
  - **Mujeres**: El gasto tiende a estar más alineado con los ingresos, lo que indica una administración financiera más estable y predecible.

2.4. Conclusiones

Una vez analizado el agrupamiento de clientes, bien de manera general y segregado por género, podemos extraer las siguientes conclusiones de forma que pueden servir como recomendación al centro comercial para maximizar sus ventas y mejorar la experiencia del cliente.

Un claro ejemplo de ello son los grupos de jóvenes, que, a pesar de tener ingresos más bajos, muestran una alta puntuación de gasto. Este entusiasmo por el consumo sugiere una oportunidad para el centro comercial de cultivar la lealtad a largo plazo mediante experiencias de compra atractivas, ofertas dirigidas y promociones en categorías de productos de alta demanda entre este segmento.

Además, los adultos mayores muestran un comportamiento más conservador en sus compras, con un gasto moderado incluso cuando sus ingresos son relativamente altos. Esto indica que pueden estar más interesados en productos y servicios que ofrezcan valor a largo plazo, como opciones de bienestar, salud y entretenimiento que justifiquen el gasto. Para captar su interés, el centro comercial podría implementar estrategias como membresías con beneficios exclusivos, programas de recompensas por compras recurrentes y eventos especializados que les brinden experiencias personalizadas.

Otro punto clave es la importancia del equilibrio entre ingresos y gasto. Mientras que algunos clientes con ingresos altos prefieren ahorrar o invertir en lugar de gastar, otros con ingresos moderados muestran una tendencia a consumir de manera más constante. Esto resalta la necesidad de diversificar la oferta del centro comercial, incluyendo tanto tiendas de lujo y productos exclusivos para clientes de alto poder adquisitivo, como opciones asequibles y promociones atractivas para quienes buscan optimizar su presupuesto sin renunciar a la calidad. Un enfoque de segmentación más detallado permitiría al centro comercial adaptar mejor su estrategia de precios y posicionamiento de marcas.

Por otro lado, la segregación por género ha sido bastante reveladora y sugiere que, para las mujeres, dado que su consumo tiende a estar más alineado con sus ingresos, es recomendable ofrecer programas de fidelización, descuentos progresivos y promociones enfocadas en su perfil de compra. Mientras que, para los hombres, existen grupos con alto poder adquisitivo pero con bajo gasto, por tanto, para incentivar el consumo, se podrían aplicar estrategias como paquetes premium, productos exclusivos o beneficios por volumen de compra, aunque también se pueden destacar opciones de inversión en experiencias de lujo (por ejemplo, tecnología, automóviles, relojería, etc.).

En definitiva, el centro comercial debe adaptar sus estrategias de venta y marketing según los hábitos de consumo de cada grupo. Mientras que las mujeres responden mejor a experiencias de compra atractivas y alineadas con sus ingresos, los hombres necesitan estrategias que resalten el valor de la inversión y la exclusividad.

Bibliografía

[1] Scikit-learn Documentation (2024). Clustering Methods.

[2] Kotler, P., Keller, K. L. (2015). Marketing Management (15th ed.). Pearson.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[4] SciPy Documentation (2024). Hierarchical Clustering (scipy.cluster.hierarchy).

[5] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[6] Raschka, S. (2015). Python Machine Learning. Packt Publishing.

Universidad de La Laguna