30 de abril del 2025

Proyecto

Perfiles que hablan: segmentación y patrones de consumo con PCA

Introducción

En un entorno empresarial cada vez más competitivo, comprender a fondo el comportamiento de los clientes se ha vuelto esencial para tomar decisiones acertadas y generar valor. Las empresas suelen recopilar una gran cantidad de datos sobre sus clientes, pero muchas veces no se aprovechan de manera estratégica. Este proyecto surge ante la necesidad de transformar esos datos en información útil que permita identificar perfiles de clientes y entender sus patrones de consumo. Utilizando herramientas de minería de datos, como el Análisis de Componentes Principales (PCA), se busca reducir la complejidad de los datos y descubrir combinaciones de variables que expliquen mejor las diferencias entre los clientes. Esto permitirá segmentarlos de forma más precisa, mejorar la toma de decisiones comerciales y anticipar el comportamiento de nuevos clientes potenciales.

Para el desarrollo de este proyecto se utilizó la base de datos Customer Personality Analysis, obtenida de la plataforma Kaggle.

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Objetivos

- Entender conceptos básicos del Análisis de Componentes Principales.
- Identificar patrones ocultos de comportamiento de consumo a través de los componentes principales.
- Interpretar visualmente la distribución de los clientes en función de sus características clave para facilitar la toma de decisiones comerciales.

Herramientas

Lenguaje de programación Python

Google Colaboratory

Contenido

1. Planteamiento del problema

La empresa cuenta con una base de datos de clientes que incluye información valiosa sobre ingresos, edad, estructura familiar y hábitos de consumo. Sin embargo, no se han identificado con claridad los diferentes tipos de clientes ni sus comportamientos de compra. Esta falta de segmentación dificulta tomar decisiones estratégicas personalizadas. Por ello, se vuelve necesario aplicar técnicas de minería de datos, como el Análisis de Componentes Principales (PCA), para descubrir patrones ocultos y definir perfiles que permitan conocer mejor a los clientes actuales y orientar acciones más efectivas hacia nuevos clientes potenciales.

2. Conociendo los datos

Visualizar las dimensiones de los datos y entenderlos.

```
data.shape
```

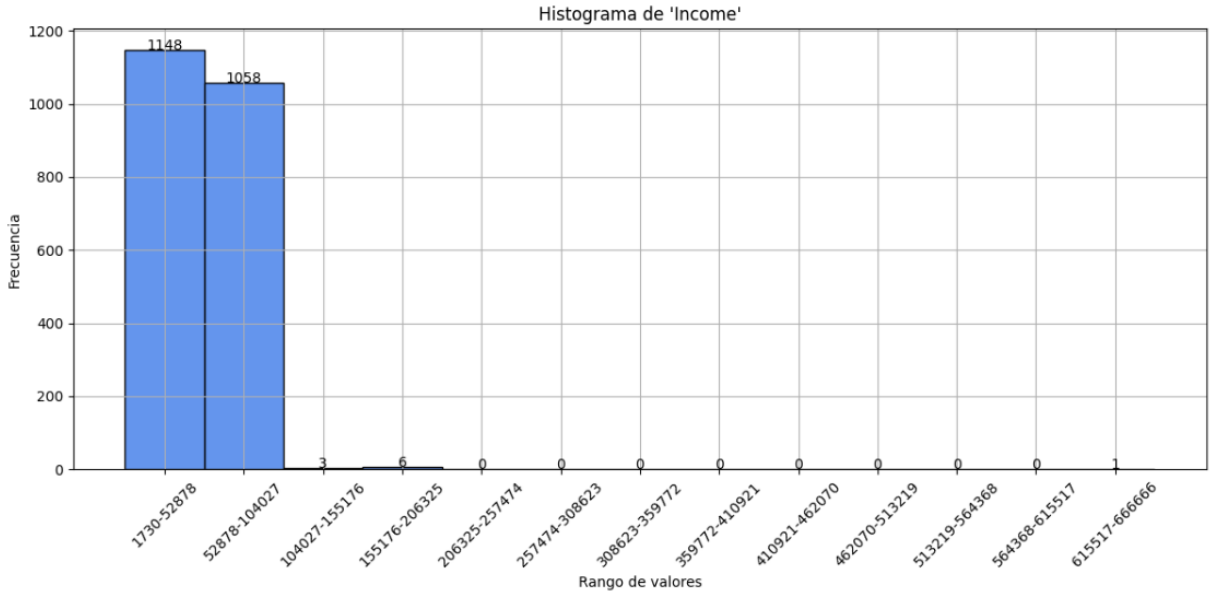
```
(2240, 28)
```

Variable	Traducción	Descripción
ID	ID (Identificador del cliente)	Identificador único del cliente
Year_Birth	Año de nacimiento	Año de nacimiento del cliente
Education	Nivel educativo	Nivel educativo del cliente
Marital_Status	Estado civil	Estado civil del cliente
Income	Ingreso anual del hogar	Ingreso anual del hogar del cliente
Kidhome	Número de niños en el hogar	Número de hijos en el hogar del cliente
Teenhome	Número de adolescentes en el hogar	Número de adolescentes en el hogar del cliente
Dt_Customer	Fecha de incorporación del cliente	Fecha en que el cliente se unió a la empresa
Recency	Días desde la última compra	Número de días desde la última compra del cliente
MntWines	Gasto en vino	Monto gastado en vino en los últimos 2 años
MntFruits	Gasto en frutas	Monto gastado en frutas en los últimos 2 años
MntMeatProducts	Gasto en carnes	Monto gastado en carne en los últimos 2 años
MntFishProducts	Gasto en pescados	Monto gastado en pescado en los últimos 2 años
MntSweetProducts	Gasto en dulces	Monto gastado en dulces en los últimos 2 años
MntGoldProds	Gasto en productos de oro	Monto gastado en productos de oro en los últimos 2 años
NumDealsPurchases	Número de compras con descuento	Número de compras realizadas con descuento
NumWebPurchases	Número de compras en la web	Número de compras realizadas a través del sitio web de la empresa
NumCatalogPurchases	Número de compras por catálogo	Número de compras realizadas mediante catálogo
NumStorePurchases	Número de compras en tienda física	Número de compras realizadas en tienda física
NumWebVisitsMonth	Visitas web por mes	Número de visitas al sitio web de la empresa en el último mes
AcceptedCmp3	Aceptó campaña 3	1 si el cliente aceptó la oferta en la campaña 3, 0 en caso contrario
AcceptedCmp4	Aceptó campaña 4	1 si el cliente aceptó la oferta en la campaña 4, 0 en caso contrario
AcceptedCmp5	Aceptó campaña 5	1 si el cliente aceptó la oferta en la campaña 5, 0 en caso contrario
AcceptedCmp1	Aceptó campaña 1	1 si el cliente aceptó la oferta en la campaña 1, 0 en caso contrario
AcceptedCmp2	Aceptó campaña 2	1 si el cliente aceptó la oferta en la campaña 2, 0 en caso contrario
Complain	Se quejó	1 si el cliente se quejó en los últimos 2 años, 0 en caso contrario
Response	Respondió última campaña	1 si el cliente aceptó la oferta en la última campaña, 0 en caso contrario

Se creó la columna “Age” (Edad) establecer la edad del cliente, ya que se cuenta con la columna “Year_Birth”.

2.1 Análisis univariado

Algunos resultados interesantes son:



1730-52878: 51.8% | 52878-104027: 47.7% | 104027-155176: 0.1% | 155176-206325: 0.3% | 206325-257474: 0.0% | 257474-308623: 0.0% | 308623-359772: 0.0% | 359772-410921: 0.0% | 410921-462070: 0.0% | 462070-513219: 0.0% | 513219-564368: 0.0% | 564368-615517: 0.0% | 615517-666666: 0.0%

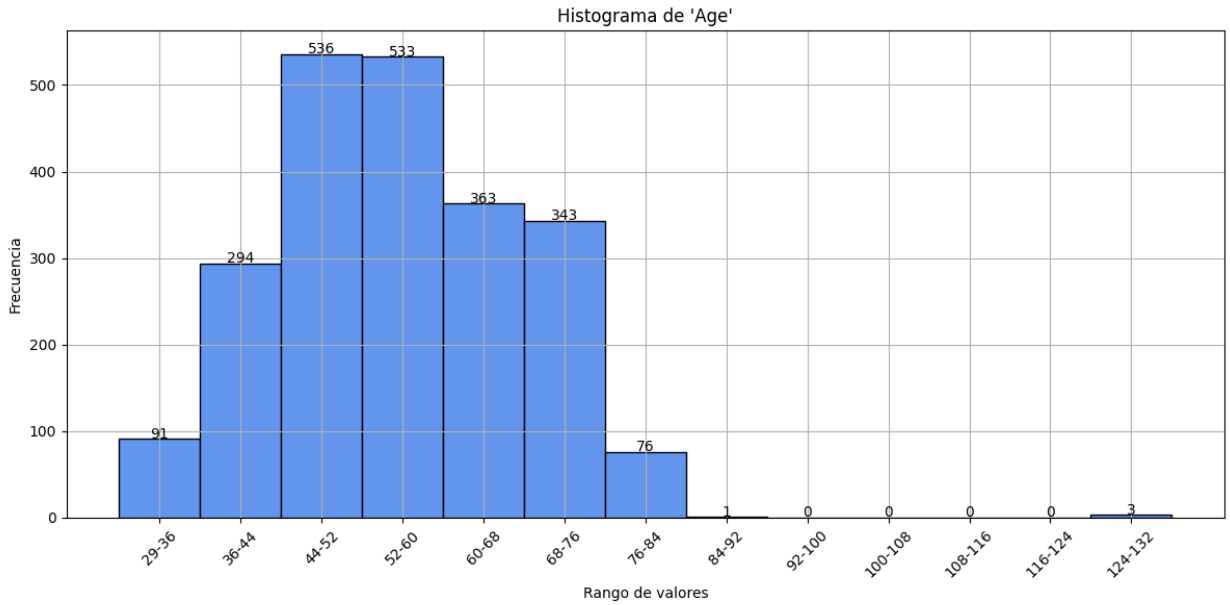
Interpretación: La gran mayoría (99%) de los ingresos anuales de los clientes se encuentran entre $1,730 y $104,027, mientras que hay 7 registros que muestran valores fuera de lo normal: 6 ingresos de $155,176 a $206,325 y otro más extremo. Dichos registros atípicos son de interés y se necesita evaluar si son casos especiales o registro por algún error.

```
Income
count     2216.00
mean     52247.25
std      25173.08
min       1730.00
25%      35303.00
50%      51381.50
75%      68522.00
max     666666.00
```

Interpretación: El resumen estadístico muestra que:

- el ingreso anual promedio es de aproximadamente $52,247, la mediana (50%) es de $51,381 y la desviación estándar de $25,173: la cercanía entre esas medidas, y aunque hay valores extremos, sugiere que la distribución no está fuertemente sesgada
- el 75% de los clientes gana hasta 68,522 al año.

- existen algunos casos con ingresos mucho más altos que el resto, como se observa en el valor máximo de 666,666, ya visto en el histograma.

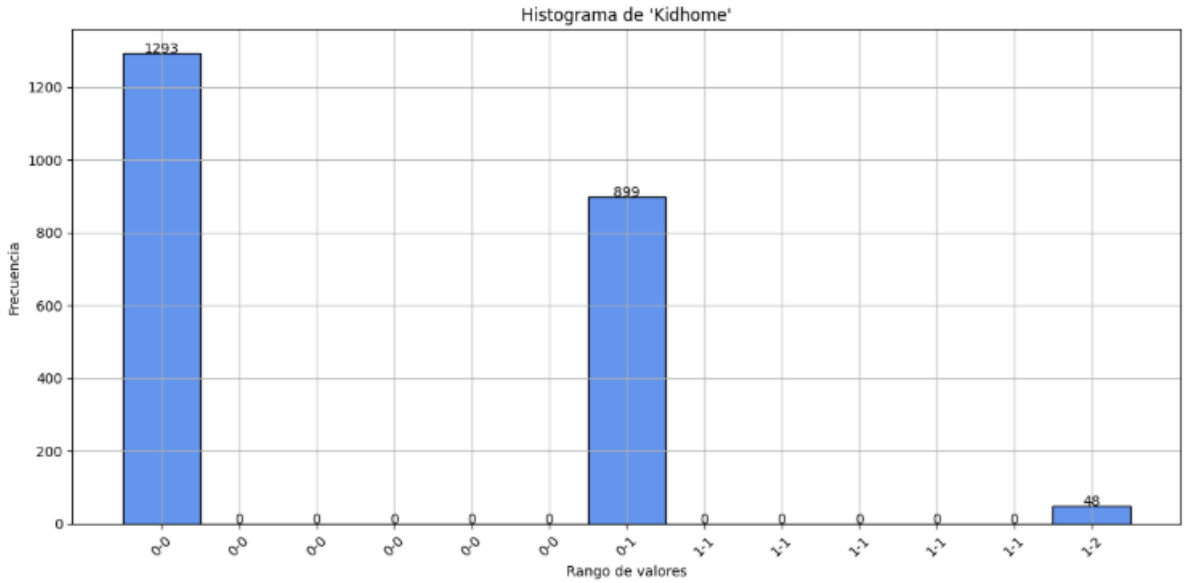


29-36: 4.1% | 36-44: 13.1% | 44-52: 23.9% | 52-60: 23.8% | 60-68: 16.2% | 68-76: 15.3% | 76-84: 3.4% | 84-92: 0.0% | 92-100: 0.0% | 100-108: 0.0% | 108-116: 0.0% | 116-124: 0.0% | 124-132: 0.1%

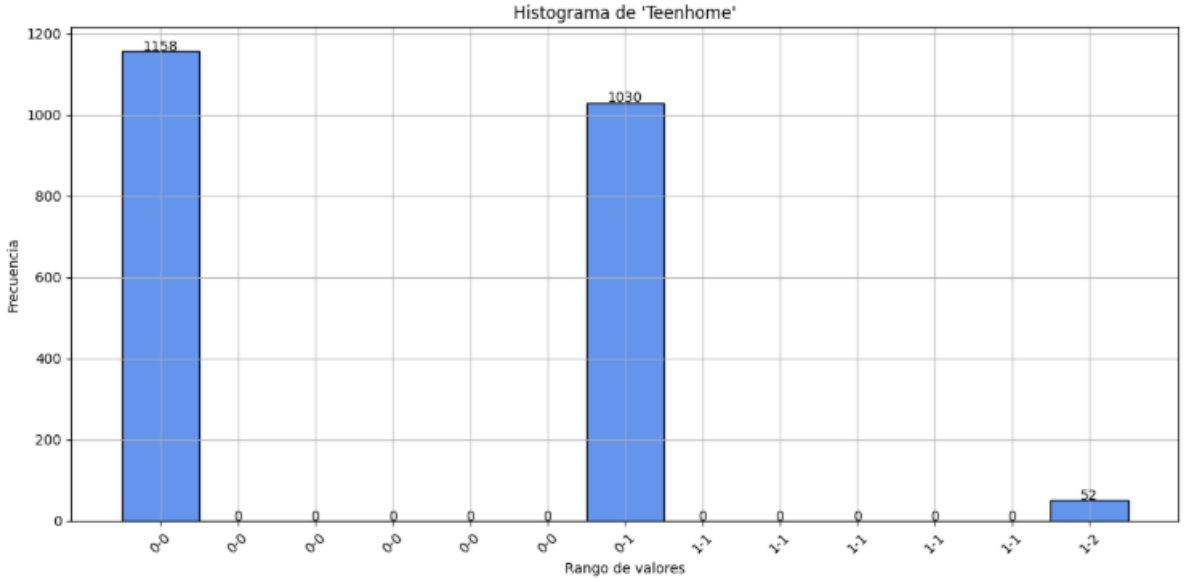
Interpretación: La gran mayoría de los clientes, casi el 96%, están en una edad madura, es decir, entre los 36 y los 76 años. Y dentro de ese grupo de personas adultas, la mitad (casi un 50%) tienen entre 44 y 60 años, lo que quiere decir que ese rango es el que más se repite. Esto nos ayuda a entender mejor a qué tipo de público se dirige la empresa, y también nos da una idea de qué edades son más comunes entre sus clientes.

Age	
count	2240.00
mean	56.19
std	11.98
min	29.00
25%	48.00
50%	55.00
75%	66.00
max	132.00

Interpretación: Los clientes tienen una edad promedio de 56 años, y la mayoría está entre los 48 y 66 años, ya que ese es el rango entre el primer y el tercer cuartil (el 50% central). La edad mínima registrada es de 29 años y la máxima llega a 132, aunque ese valor parece un dato extraño o error de registro. Además, como la mediana (55) está muy cerca del promedio (56.19), podemos decir que la distribución no está muy sesgada, y que la mayoría de los datos están bien centrados.



0-0: 57.7% | 0-0: 0.0% | 0-0: 0.0% | 0-0: 0.0% | 0-0: 0.0% | 0-0: 0.0% | 0-1: 40.1% | 1-1: 0.0% | 1-1: 0.0% | 1-1: 0.0% | 1-1: 0.0% | 1-1: 0.0% | 1-2: 2.1%

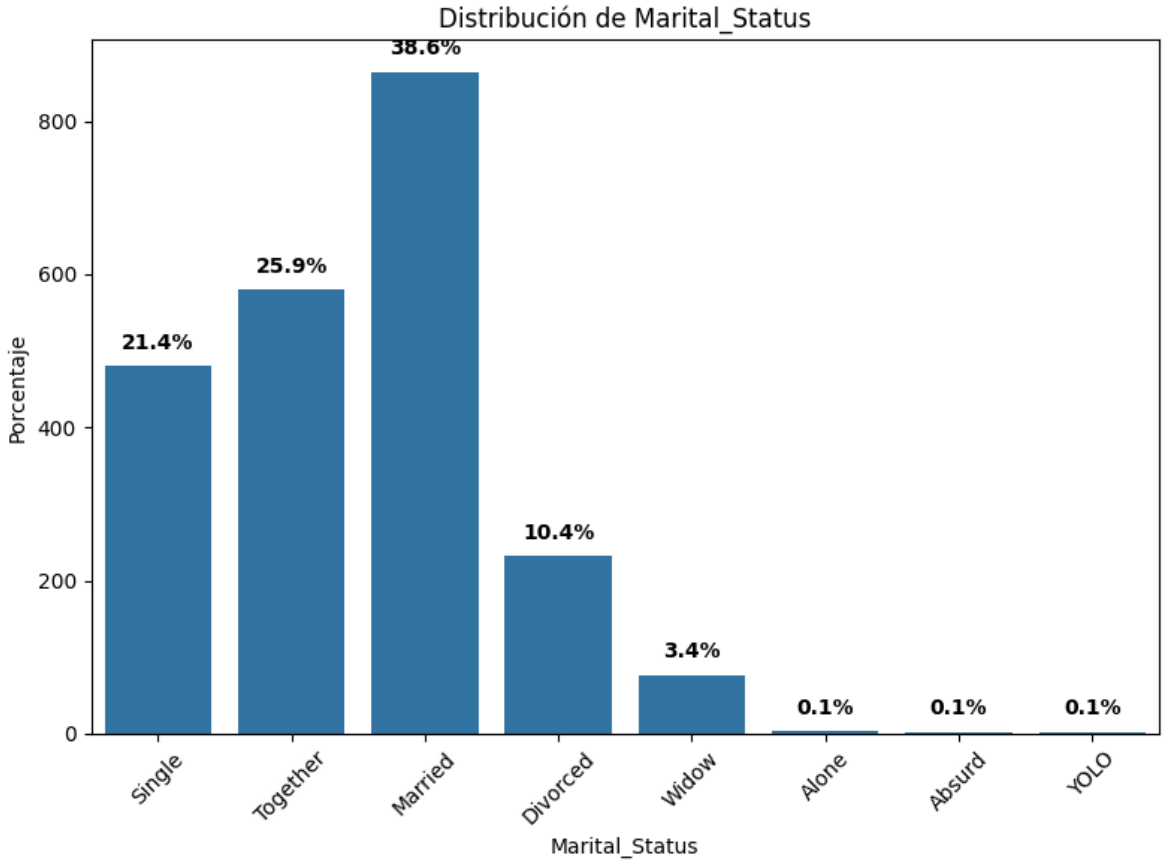


0-0: 51.7% | 0-0: 0.0% | 0-0: 0.0% | 0-0: 0.0% | 0-0: 0.0% | 0-0: 0.0% | 0-1: 46.0% | 1-1: 0.0% | 1-1: 0.0% | 1-1: 0.0% | 1-1: 0.0% | 1-1: 0.0% | 1-2: 2.3%

Interpretación: La mayoría de los clientes no tienen niños pequeños en casa: un 57.7% no tiene hijos y un 40.1% tiene solo 1 hijo en casa. Muy pocos tienen 2 hijos (apenas un 2.1%).

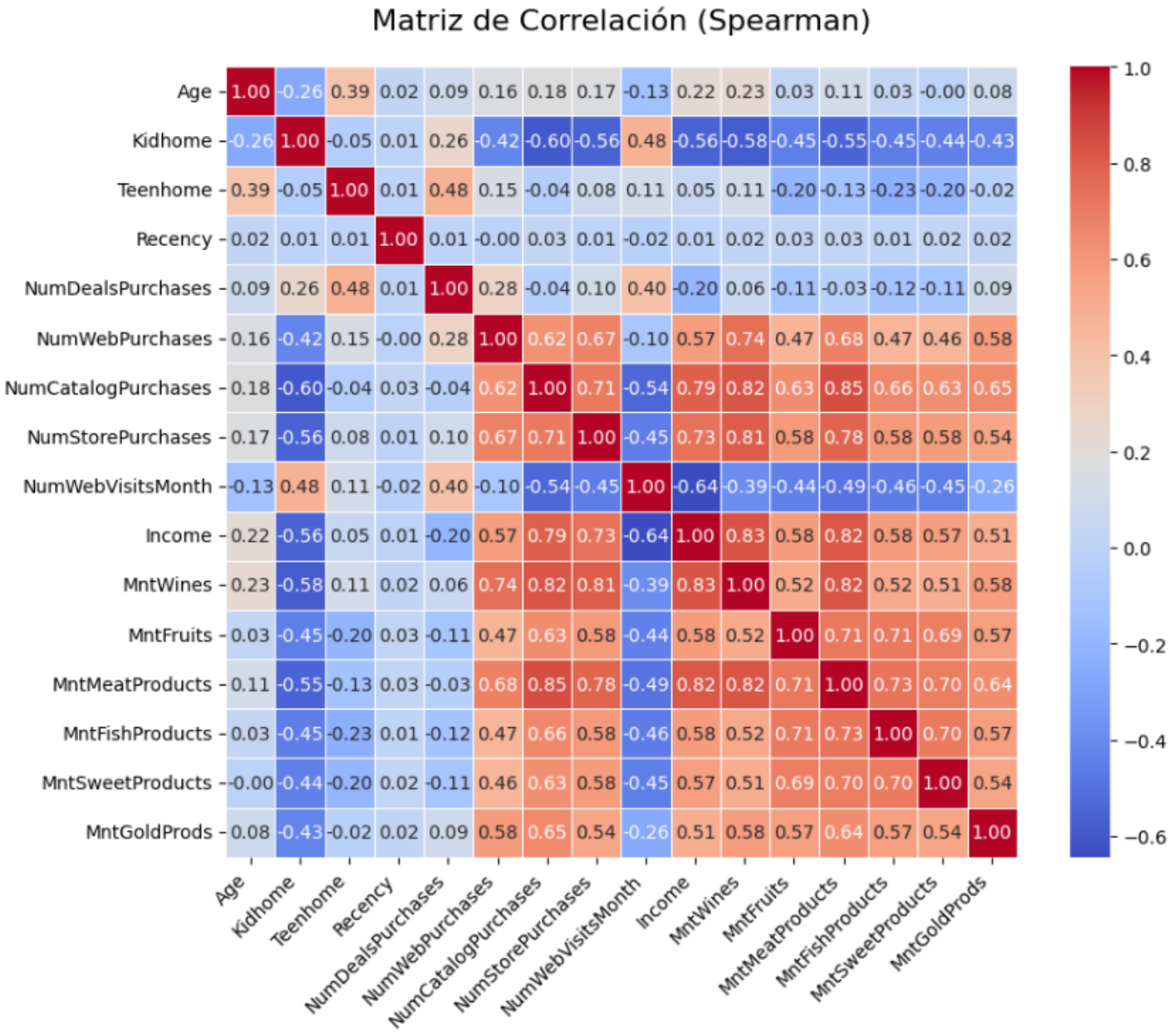
En cuanto a adolescentes (Teenhome), el patrón es muy similar: el 51.7% no tiene adolescentes en el hogar, mientras que 46% tiene uno solo, y apenas un 2.3% tiene dos adolescentes.

Lo que puede tener relación con los rangos de edad de los clientes, teniendo hijos ya independientes, es posible que tengan más capacidad de inversión personal.

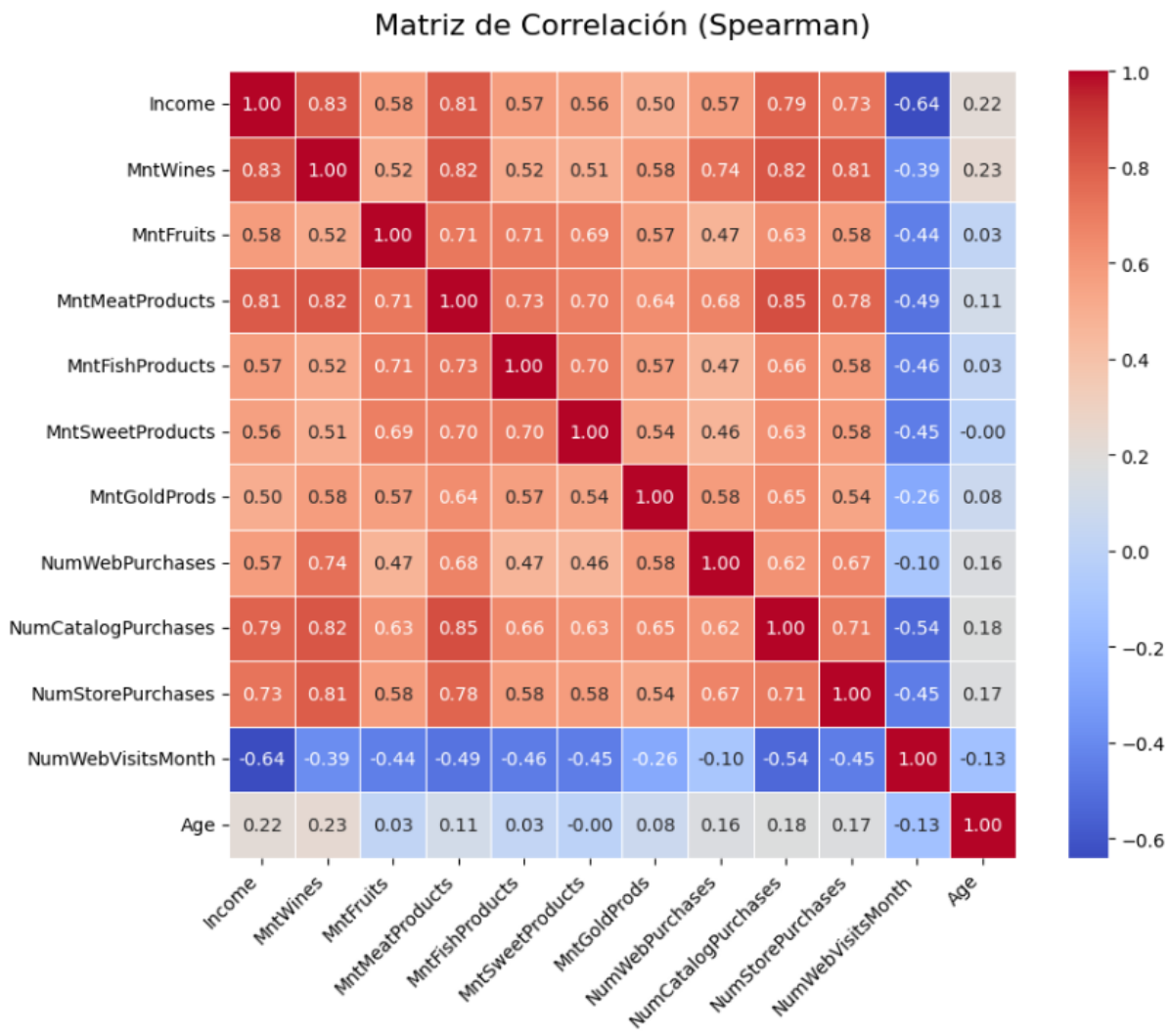


Interpretación: Se puede observar que, aproximadamente, el 65% de los clientes tienen parejas, ya en matrimonio (38.6%) o viviendo juntos (25.9%). También es notable que un 21% afirma estar solteros. Se puede suponer que la gran mayoría de los clientes tienen gastos compartidos, mientras que un 33% puede más independencia personal.

2.2 Análisis Bi variado



Interpretación: Se conservarán las variables con más alta correlación que son: Income, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds y se determina, por **criterio de experto**, descartar las otras variables.



2.3 Evaluación de datos nulos

No se detectaron datos nulos

2.4 Tratamiento de datos atípicos

Se aplicó winsorización a los valores atípicos usando los percentiles 10 y 90.

3. Modelado

El Análisis de Componentes Principales es una técnica estadística que permite reducir la cantidad de variables en un conjunto de datos, combinándolas en nuevas variables llamadas "componentes principales". Estos componentes capturan la mayor parte de la información de los datos originales, pero de manera más compacta ayudando a simplificar el análisis y a descubrir relaciones interesantes entre los datos que de otra forma pasarían desapercibidas.

Esta técnica trabaja con variables numéricas continuas y discretas correlacionadas significativamente. Antes de aplicarla es importante estandarizar los datos.

Parte importante de la interpretación de los resultados del análisis de componentes principales son los valores y vectores propios.

3.1 Valores propios

Son números que indican cuánta información (o cuánta "importancia") tiene cada componente principal.

Un valor propio grande = ese componente guarda mucha información de los datos.

Un valor propio pequeño = ese componente guarda poca información.

3.2 Vectores propios

Son las combinaciones lineales de las variables originales y constan de un signo (positivo o negativo) y el peso o importancia de la variable original para el componente.

4. Resultados

Se determina armar 3 componentes principales cuyos resultados fueron:

Valores propios:

```
[64.2314408  10.49782484  9.13334008]
```

- El primer valor propio es 64.23 → el primer componente principal guarda muchísima información.
- El segundo valor propio es 10.50 → el segundo componente principal guarda lo que el primer componente no pudo capturar, es menos información, pero todavía es importante.
- El tercer valor propio es 9.13 → el tercer componente principal guarda lo que el segundo componente no pudo capturar, es un poquito menos que el segundo, pero sigue siendo útil.

En total, los tres componentes principales capturan un aproximado del 83.86% de la información de los datos originales.

Vectores propios

	Income	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
PC1	0.40	0.36	0.38	0.42	0.39	0.38	0.31
PC2	0.40	0.65	-0.36	0.11	-0.36	-0.37	-0.06
PC3	-0.16	0.00	-0.10	-0.19	-0.10	-0.19	0.94

Basados en el aporte de las variables originales y sus signos, se puede determinar dos tipos de clientes para cada uno de los componentes:

Primer componente (PC1):

- Clientes con buenos ingresos y consumo equilibrado en vinos, carnes y otros productos.
- Clientes con ingresos moderados y bajo consumo general en los productos.

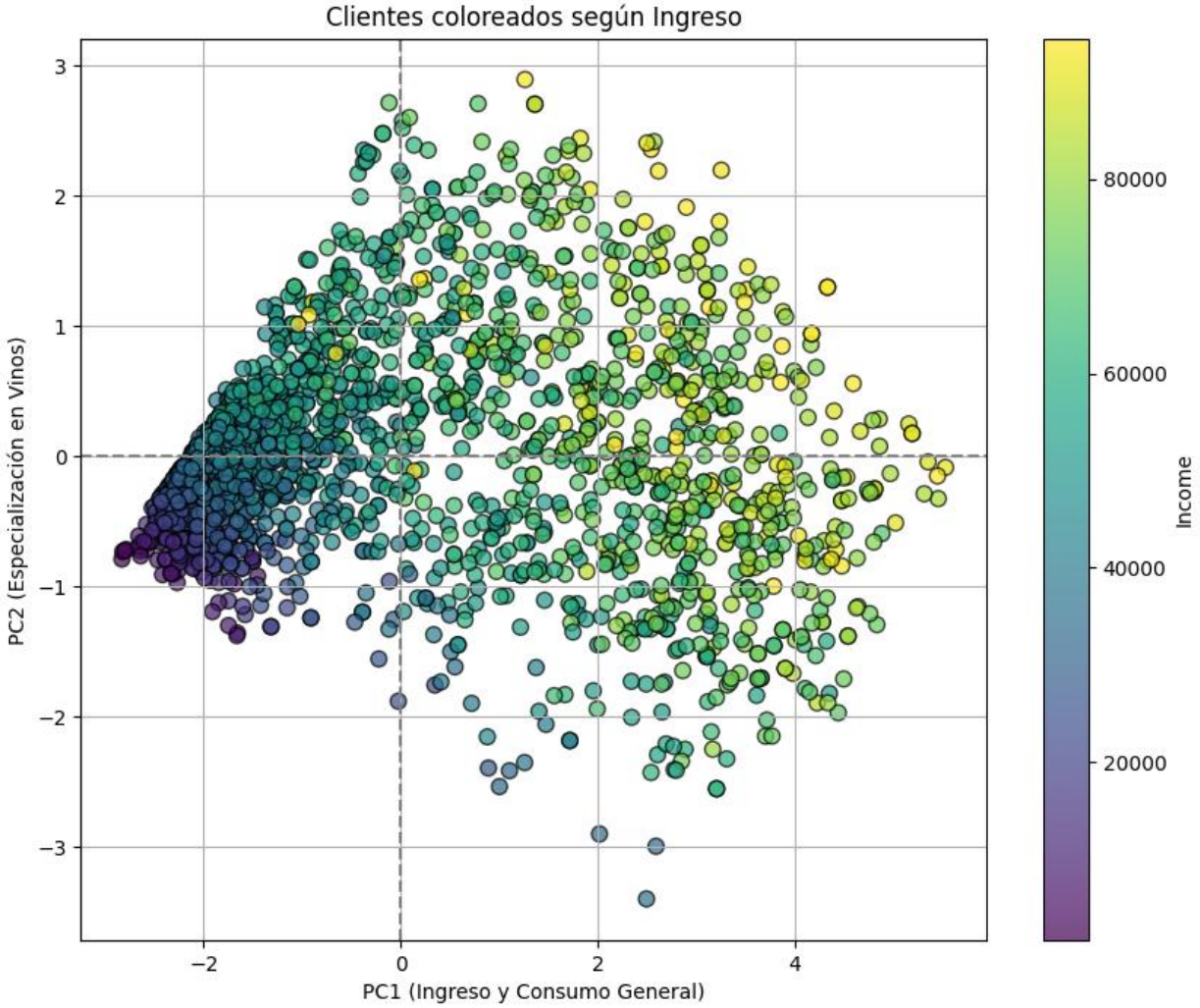
Segundo componente (PC2):

- Clientes que prefieren carnes y sobre todo vino.
- Clientes que prefieren frutas, pescados y dulces.

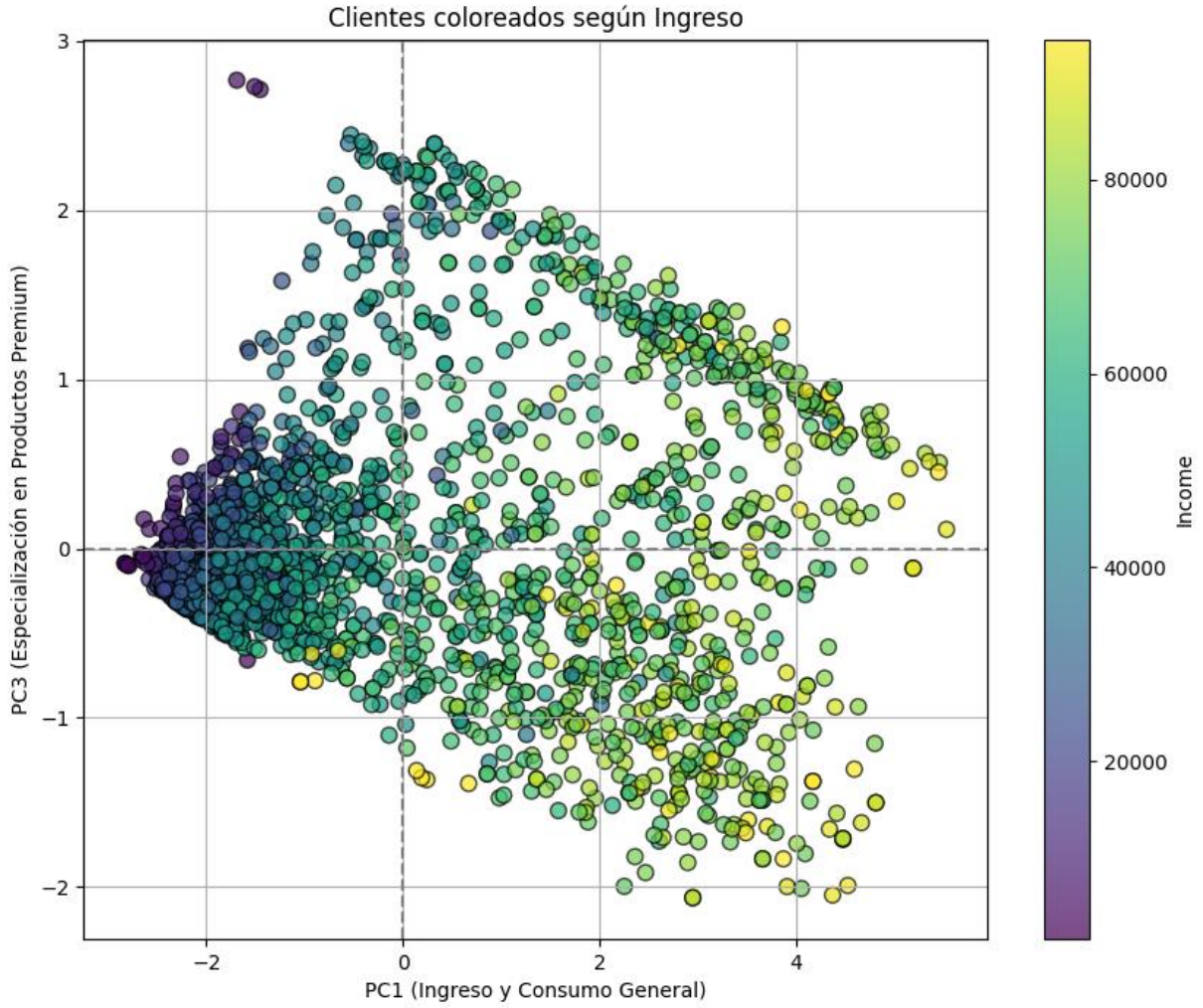
Tercer componente (PC3):

- Clientes con ingresos moderados que compran muchos productos premium (Gold).
- Clientes con buenos ingresos que prefieren productos más básicos o tradicionales.

4.1 Gráficos



Interpretación: La mayoría de los clientes muestra un comportamiento organizado: los clientes con mayores ingresos (colores más claros) tienden a concentrarse hacia la derecha del gráfico, donde se observa un mayor consumo general, mientras que los clientes con menores ingresos (colores más oscuros) se agrupan hacia la izquierda, con un consumo más bajo. Además, la especialización en vinos (eje vertical) no presenta grandes extremos para la mayoría de los clientes, quienes mantienen un comportamiento de consumo moderado.



Interpretación: En general, la mayoría de los clientes muestra un comportamiento relativamente agrupado: los clientes con mayores ingresos tienden a concentrarse hacia la derecha del gráfico, con un consumo general más alto, mientras que la especialización en productos premium se mantiene moderada para la mayoría, sin grandes extremos.

Sin embargo, se observa un pequeño grupo de clientes con ingresos bajos que se posicionan muy alejados del resto en el eje de especialización en productos premium. Esto puede indicar la presencia de clientes con comportamientos de compra muy inusuales o posibles valores atípicos en los datos, que sería importante revisar con más detalle.

5. Conclusiones

- Conocer a nuestros clientes a través de sus datos reales permite tomar decisiones más inteligentes. A partir del análisis de variables como ingresos, edad y hábitos de consumo, se identificaron características comunes que describen claramente a los clientes actuales de la empresa.
- El uso del Análisis de Componentes Principales (PCA) como modelo no supervisado permitió descubrir patrones ocultos en los datos.
- Al conocer cómo se comportan distintos grupos de clientes, es posible diseñar estrategias personalizadas, mejorar la segmentación de mercado, y enfocar esfuerzos en atraer o fidelizar a determinados perfiles.
- Algunos clientes mostraron comportamientos inusuales, como alta preferencia por productos premium a pesar de bajos ingresos. Estas excepciones pueden ser oportunidades o advertencias para revisar la calidad de los datos o explorar nuevos nichos de mercado.
- Los comportamientos pasados y actuales de los clientes permiten proyectar cómo podrían actuar nuevos clientes, mejorar la toma de decisiones comerciales y anticiparse a cambios en las preferencias.