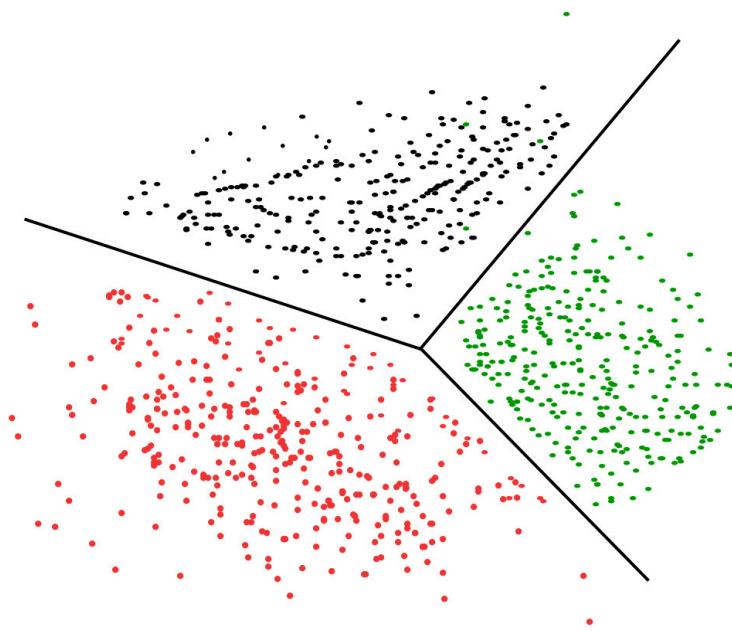


# Clustering E-Commerce

*Aprenentatge Computacional*



Martí Gasol, Marc Jarauta, Lucas Carbó, Santi Prats

Grup 04

Grau en Enginyeria de Dades

Curs 2025-2026

# 1. Introducció

L'objectiu d'aquest informe és descriure el procés d'exploració, neteja i anàlisi preliminar d'un conjunt de dades compost per 2.240 registres de clients i 29 atributs. Aquest dataset inclou variables numèriques, categòriques i temporals relacionades amb característiques demogràfiques, patrons de consum, respostes a campanyes promocionals i canals de compra.

L'estat inicial de les dades presenta inconsistències, valors atípics i informació incompleta, motiu pel qual s'ha realitzat un treball exhaustiu de depuració per garantir la qualitat abans de procedir a l'aplicació d'algorismes de clustering.

## 2. Descripció del Dataset Original

### 2.1 Estructura General

El dataset combina tres tipus principals de variables:

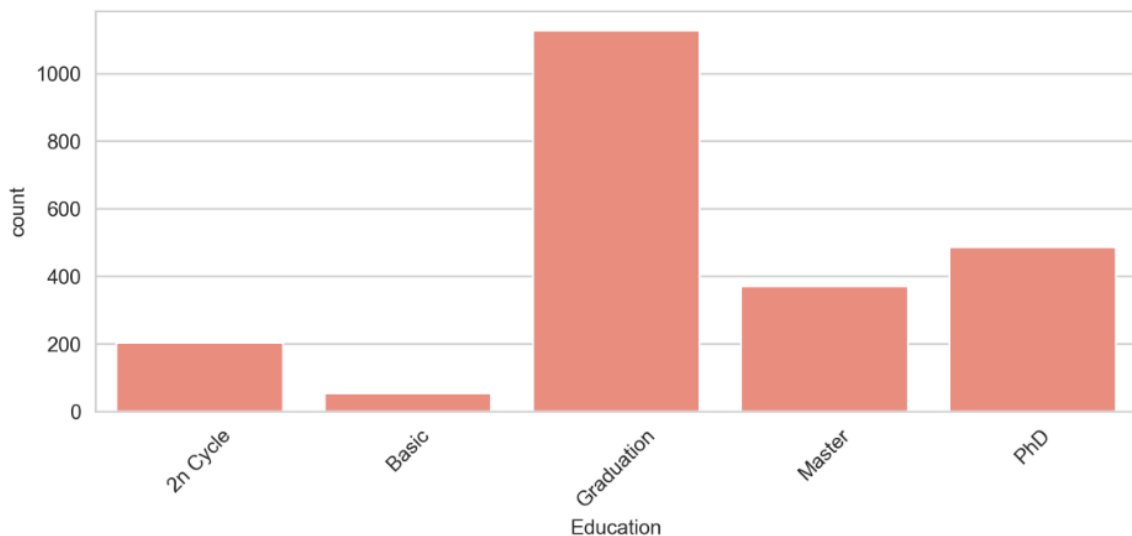
- **Variables numèriques:** ingressos, despeses per categoria de producte, nombre de visites web, any de naixement, etc.
- **Variables categòriques:** nivell educatiu i estat civil.
- **Variables temporals:** data d'alta del client, utilitzada eventualment per calcular l'antiguitat.

Aquest conjunt de dades és ric i detallat, però inicialment presentava un elevat nivell de soroll i inconsistències que requerien un procés acurat de neteja.

## 2.2 Nivell Educatiu

La variable Education és de naturalesa categòrica ordinal i descriu el grau acadèmic assolit pel client. En analitzar la distribució inicial de les dades, hem observat que la mostra presenta cinc nivells distints; Basic, 2n Cycle, Graduation, Master i PhD.

Com es pot apreciar en l'anàlisi gràfica, la distribució està desbalancejada: la categoria "Graduation" representa el grup majoritari amb més de 1.000 clients, mentre que nivells com "Basic" i "2n Cycle" tenen una representació molt minoritària.



Atès que els algorismes de clustering basats en distància (com K-Means) requereixen entrades estrictament numèriques, no podem utilitzar les etiquetes de text directament. A més, existeix una jerarquia lògica en l'educació (un Doctorat és un grau superior a un Màster, que és superior a un Grau, etc.).

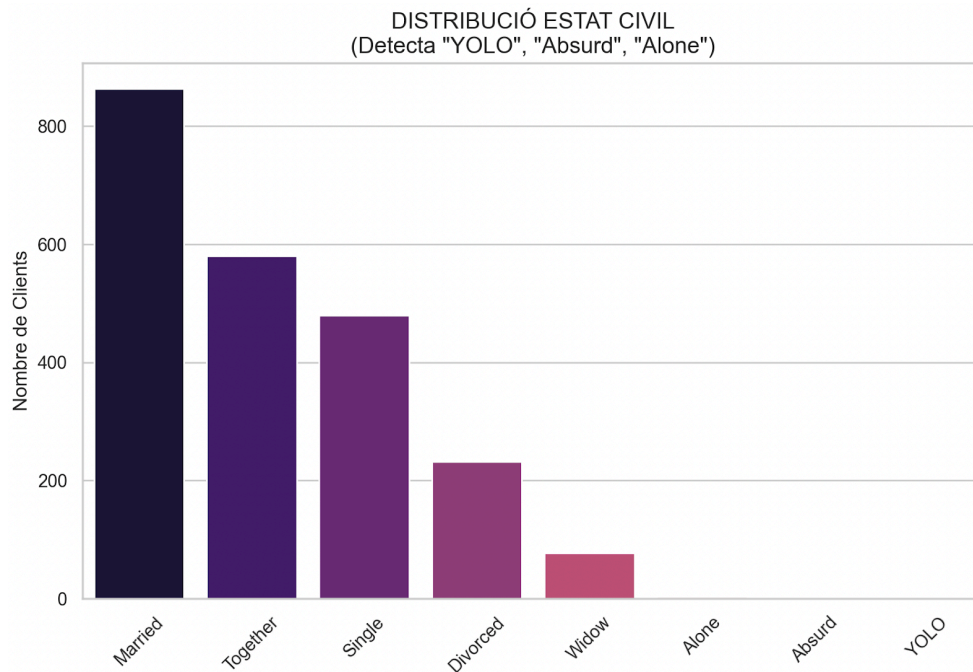
Per respectar aquesta jerarquia i no perdre informació valuosa durant la transformació, hem aplicat una Codificació Ordinal. En lloc de tractar-les com a categories independents, les hem mapejat en una escala lineal de l'1 al 5, on un valor més alt indica un nivell d'estudis superior.

El criteri de codificació aplicat ha estat el següent:

```
# Educació
education_map = {'Basic': 1, '2n Cycle': 2, 'Graduation': 3, 'Master': 4, 'PhD': 5}
data['Education_Code'] = data['Education'].map(education_map).fillna(0)
```

## 2.3 Estat Civil

La variable `Marital_Status` contenia valors no vàlids com “*Alone*”, “*Absurd*” i “*YOLO*”, que no corresponen a cap estat civil real. Aquests registres s’han eliminat, ja que no aportaven informació coherent i podrien distorsionar els resultats del clustering.



Un cop netejada la variable, hem observat una fragmentació excessiva en les categories restants ("Single", "Divorced", "Widow", "Married", "Together"). Des d'una perspectiva de comportament de compra i estructura de la llar, un client divorciat o vidu comparteix característiques econòmiques molt similars a les d'un client solter (tots tres formen, generalment, nuclis monoparentals o unipersonals).

Per simplificar la dimensionalitat del model sense perdre informació rellevant, hem optat per una estratègia d'agrupació lògica. Hem unificat els estats de "solteria" en una única categoria i hem assignat codis numèrics (*Label Encoding*) per distingir els diferents graus de convivència en parella.

El mapeig final aplicat ha estat el següent:

```
# Estat Civil (Agrupant Divorced/Widow/Single com a 3)
marital_status_map = {'Married': 1, 'Together': 2, 'Single': 3, 'Divorced': 3, 'Widow': 3}
data['Marital_Status_Code'] = data['Marital_Status'].map(marital_status_map).fillna(0)
```

## 2.4 Ingressos

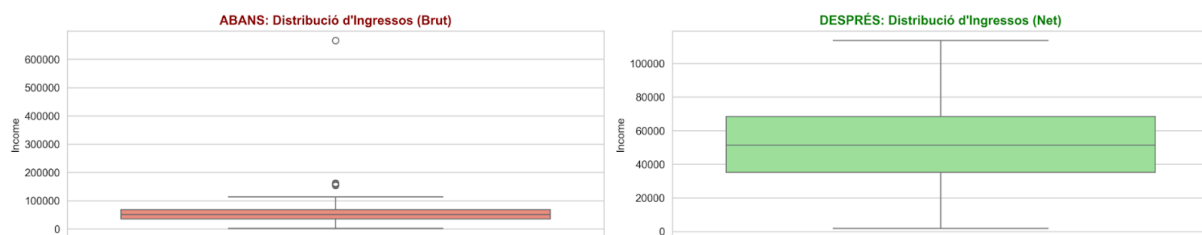
En l'exploració inicial, hem detectat l'absència d'informació financera en 24 registres (aproximadament l'1% de la mostra). Atès que la imputació artificial d'ingressos (mitjançant mitjanes o medianes) podria introduir un biaix considerable en un factor tan sensible, hem optat per eliminar aquests registres.

També vam detectar la presència d'un valor extremadament atípic de 666.666 €, que també vam eliminar.

Fins i tot després d'eliminar el cas extrem, l'anàlisi gràfica mostrava encara una sèrie de valors "flotants" a la franja superior (ingressos entre 100.000 € i 160.000 €) que es desviaven significativament de la tendència central del dataset.

Per homogeneïtzar la mostra i centrar l'estratègia de negoci en el comportament del client representatiu (evitant que el model es veiés esbiaixat per una minoria d'alt poder adquisitiu), hem aplicat el criteri del Rang Interquartílic (IQR):

1. S'han calculat el primer (Q1) i tercer quartil (Q3) de la distribució.
2. S'ha definit un llindar superior estricte mitjançant la fórmula:  $\text{Límit} = Q3 + 1.5 \times \text{IQR}$ .
3. Tots els registres que superaven aquest llindar matemàtic han estat filtrats del dataset final



## 2.5 Variables de Comportament i Canal

Tot i que aquest conjunt de variables és fonamental per a la comprensió del negoci, hem pres la decisió estratègica d'excloure-les de la matriu d'entrenament del model (K-Means i PCA). Aquestes variables no actuen com a *Variables Actives* (les que creen els grups), sinó com a *Variables Il·lustratives* (les que ens ajuden a explicar-los a posteriori). Els motius tècnics d'aquesta exclusió són els següents:

- **Despesa per Categoria (MntWines, MntMeatProducts, etc.)** Aquestes variables presenten un problema de multicolinealitat directa amb la variable Total\_Spending que hem creat. Com que la despesa total és la suma lineal d'aquestes categories, incloure-les totes dues duplicaria el pes del factor "diners" dins l'algorisme, fent que el model ignorés altres

dimensions com l'antiguitat o l'estructura familiar. Hem prioritzat segmentar primer per Poder Adquisitiu(Total) i analitzar les preferències de producte un cop definits els grups.

**-Canals de Compra (NumWebPurchases, etc.)** La inclusió d'aquestes variables tendia a introduir soroll en la segmentació. L'objectiu del clustering és identificar el *valor* i el *cicle de vida* del client, no tant el seu mètode d'accés. Un client VIP ho és tant si compra per web com per catàleg. Incloure els canals a l'inici fragmentava els clústers innecessàriament. Aquestes mètriques s'utilitzen, en canvi, per definir l'estratègia de comunicació de cada segment resultant.

**-Resposta a Promocions (AcceptedCmp...)** Aquestes variables són de naturalesa binària (0/1) i presenten una alta dispersió (sparsity), ja que la majoria de clients no han acceptat campanyes prèvies. Barrejar variables binàries amb variables contínues d'alta magnitud (com els Ingressos) en un model basat en distància Euclidiana (K-Means) sol reduir la qualitat de la silueta dels clústers. Per tant, s'utilitzen exclusivament com a indicador de rendiment (KPI) per mesurar l'èxit potencial de les accions proposades per a cada grup.

## 2.6 Altres

**-Any de Naixement (Year\_Birth):** S'han detectat dates de naixement anteriors a 1920 (implicant edats superiors a 100 anys). Aquests casos s'han tractat com a errors de dades i s'han eliminat.

**-Composició de la Llar (Kidhome, Teenhome):** Indiquen el nombre de nens petits i adolescents a la llar, variables clau per entendre les necessitats de consum familiar.

**-Data d'Alta (Dt\_Customer):** Registra el moment en què el client es va inscriure a l'empresa. Aquesta variable no s'utilitza en el seu format original (data), sinó que serveix de base per calcular la nova variable d'**Antiguitat** (*Seniority*) que descriurem més endavant.

**-Recència (Recency):** Indica el nombre de dies transcorreguts des de l'última compra. Tot i ser una mètrica habitual en models RFM, en aquest cas s'utilitza com a variable de control post-segmentació per identificar clients "adormits" dins de cada clúster.

**-Queixes (Complain):** Variable binària que assenyala si el client ha presentat alguna reclamació en els darrers dos anys. Atès que la immensa majoria de registres tenen valor 0 (baixa variància), s'ha optat per no incloure-la al nucli del model de clustering per evitar generar grups artificials basats només en un fet puntual negatiu.

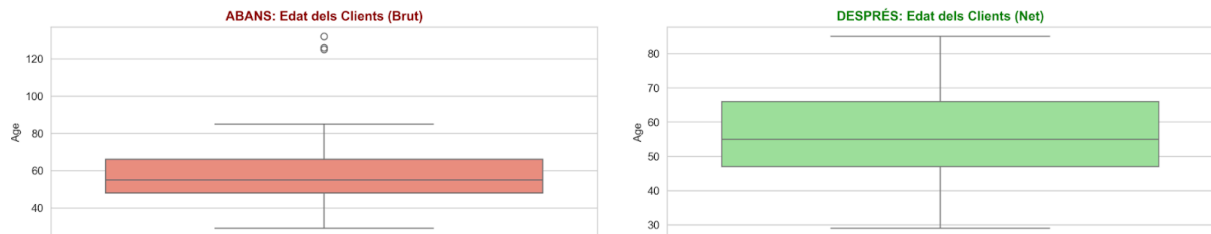
## 3. Noves Variables

### 3.1 Edat

Transformem l'any de naixement a l'edat real del client per facilitar la interpretació.

-Càlcul:  $2025 - \text{Year\_Birth}$ .

Com s'ha esmentat anteriorment, hem filtrat edats superiors a 100 anys per eliminar dades errònies.

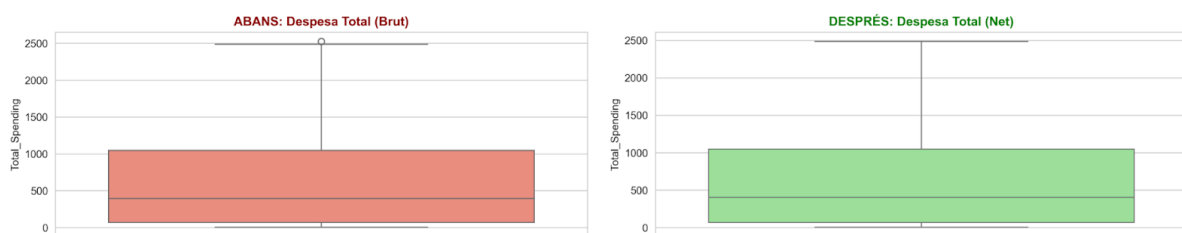


### 3.2 Despesa Total (Total\_Spending)

Representa el valor monetari total del client, simplificant les 6 categories de producte en un únic indicador de valor.

-Càlcul: Sumatòri de totes les variables Mnt....

Hem detectat algun client amb una despesa extremadament alta que podrien esbiaixar els clústers. Hem aplicat el mètode del Rang Interquartílic (IQR) per identificar i eliminar els que se situaven molt per sobre del comportament habitual.



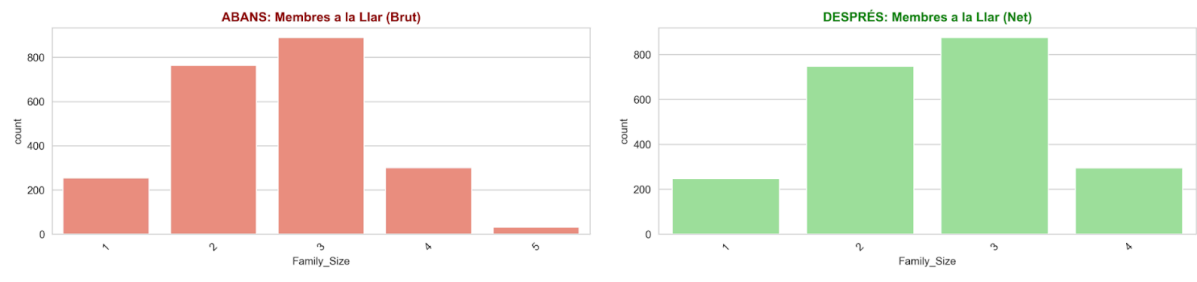
### 3.3 Mida Familiar (Family\_Size)

Variable que sintetitza tota la informació de la llar en un sol número (nombre de persones).

-Càlcul: Hem creat una variable binària Has\_Partner (1 si està casat/junt, 0 si no) i hem sumat:  $1(\text{Client}) + \text{Has\_Partner} + \text{Kidhome} + \text{Teenhome}$ .

```
partner_status = ['Married', 'Together']
data['Has_Partner'] = data['Marital_Status'].apply(lambda x: 1 if x in partner_status else 0)
data['Family_Size'] = 1 + data['Has_Partner'] + data['Kidhome'] + data['Teenhome']
```

Hem aplicat novament el criteri IQR per detectar famílies amb una mida inusualment gran (errors de dades o casos extremadament rars) i les hem filtrat per evitar soroll en la segmentació.



### 3.4 Antiguitat (Seniority i Tenure\_Days)

Mesura de la fidelitat i el temps de relació amb l'empresa.

-Càlcul: Hem convertit la data de registre (Dt\_Customer) a dies transcorreguts fins a l'actualitat (Tenure\_Days).

Per facilitar la segmentació, hem codificat aquesta variable contínua en dues categories clares: Recent (nous usuaris), i Senior (usuaris veterans), creant així la variable Seniority\_Code.

```
data['Seniority'] = pd.cut(
    data['Tenure_Days'],
    bins=[-np.inf, 365, np.inf],
    labels=['Recent', 'Senior']
)
data['Seniority_Code'] = data['Seniority'].map({'Recent': 1, 'Senior': 2})
```

## 4. Normalització i Escalat de Dades

Un pas crític abans d'aplicar qualsevol algoritme de clustering basat en distàncies és l'escalat de les dades. El nostre dataset conté variables amb magnituds extremadament dispars: mentre que els Ingressos es mouen en rangs de desenes de milers, altres variables com Family\_Size o els codis d'Educació es mouen en unitats d'una sola xifra (1 a 5).

Si introduïssim les dades brutes al model, la variable d'Ingressos dominaria completament el càlcul de la distància Euclidiana, fent que la resta de variables fossin irrelevantes. Per evitar aquest biaix i assegurar que totes les característiques contribueixin equitativament a la formació dels segments, hem avaluat les dues tècniques d'escalat més robustes i esteses en la literatura de ciència de dades:

### 4.1 Tècniques Avaluades

- **Min-Max Scaler (Normalització):** Aquesta tècnica transforma totes les variables perquè s'ajustin estrictament a un rang tancat entre 0 i 1. La fórmula aplicada és:

$$\frac{x - \min}{\max - \min}$$

El principal avantatge que hem considerat és que preserva exactament la forma de la distribució original de les dades i és molt intuïtiu quan es treballa amb variables categòriques codificades (com els nostres codis d'1 a 5), ja que simplement les "comprimeix" sense distorsionar les distàncies relatives entre categories.

- **Standard Scaler (Estandardització):** Aquesta tècnica centra les dades al voltant del 0 i les escala segons la seva desviació estàndard (Variance Scaling). La fórmula és:

$$\frac{x - \text{mitjana}}{\text{desviació estàndard}}$$

Aquest mètode és ideal quan les dades segueixen una distribució Normal (Gaussiana). No obstant això, no fixa uns límits màxims o mínims, per la qual cosa les variables amb una variància molt alta poden continuar tenint un pes superior a les altres.

### 4.2 Selecció Final de l'Escalador

Per determinar quina tècnica oferia una millor qualitat de segmentació per al nostre cas específic, hem realitzat una prova empírica comparant el rendiment de tots dos mètodes sobre el mateix conjunt de dades net.

Com es podrà observar detalladament a la taula comparativa del Punt 8, els resultats han estat concloents a favor del Min-Max Scaler.

Hem observat que l'Standard Scaler penalitzava excessivament les variables discretes (com l'Antiguitat o l'Educació) en tenir aquestes una desviació estàndard molt menor comparada amb la dels Ingressos o la Despesa. Això provocava clústers menys definits i amb mètriques

---

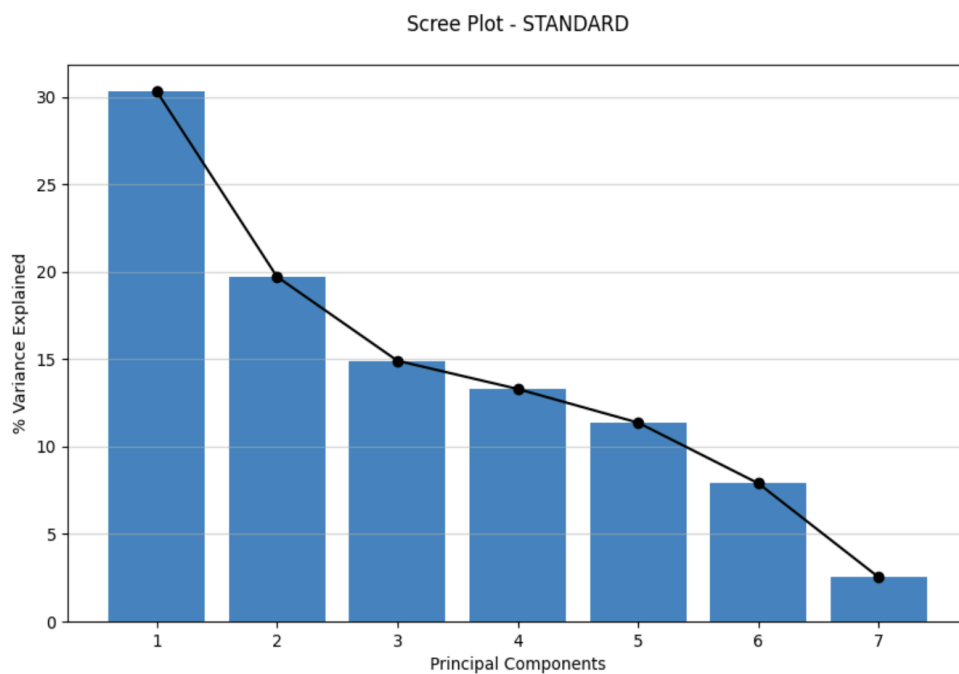
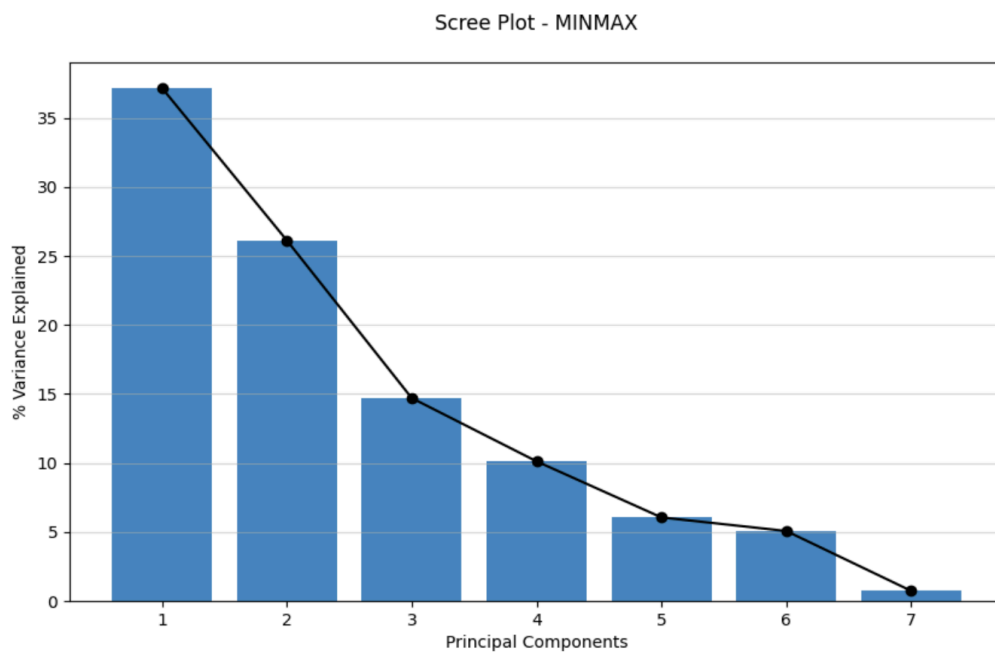
de cohesió més pobres. En canvi, el Min-Max Scaler ha permès democratitzar el pes de totes les variables, aconseguint que la diferència entre ser "Nou" o "Antic" (0 vs 1) sigui tan important per al model com la diferència entre tenir ingressos alts o baixos.

Per tant, tots els resultats de clustering presentats en aquest informe s'han generat utilitzant dades normalitzades amb Min-Max Scaler.

## 5. Tria de Variables

### 5.1 Tria de Components pel PCA

Per decidir quines variables usar al clustering hem utilitzat el PCA, aquest mètode ens dona una llista de components amb la seva variància explicada. Ens quedem amb tantes components com siguin necessàries per reunir el 80% (Valor òptim per explicar la major part de les dades però no tant com per tenir massa soroll i complicar el model)



A partir dels dos gràfics hem pogut calcular que en amndos casos és a partir de la quarta component quan superem aquest 80% de la variància explicada.

## 5.2 Tria de Variables per a cada cas

```

=====
PCA Component Loadings (minmax):
=====

PC1:
-----
Seniority_Code      0.973830
Marital_Status_Code 0.165375
Total_Spending      0.109791
Family_Size         0.105059
Education_Code       0.025143
Income              0.023378
Name: PC1, dtype: float64

PC2:
-----
Marital_Status_Code 0.872836
Family_Size         0.430339
Seniority_Code       0.205339
Total_Spending       0.083669
Income              0.059879
Age                 0.011066
Name: PC2, dtype: float64

PC3:
-----
Total_Spending      0.670879
Income              0.490724
Family_Size         0.373820
Marital_Status_Code 0.303401
Education_Code       0.226416
Age                 0.144819
Name: PC3, dtype: float64

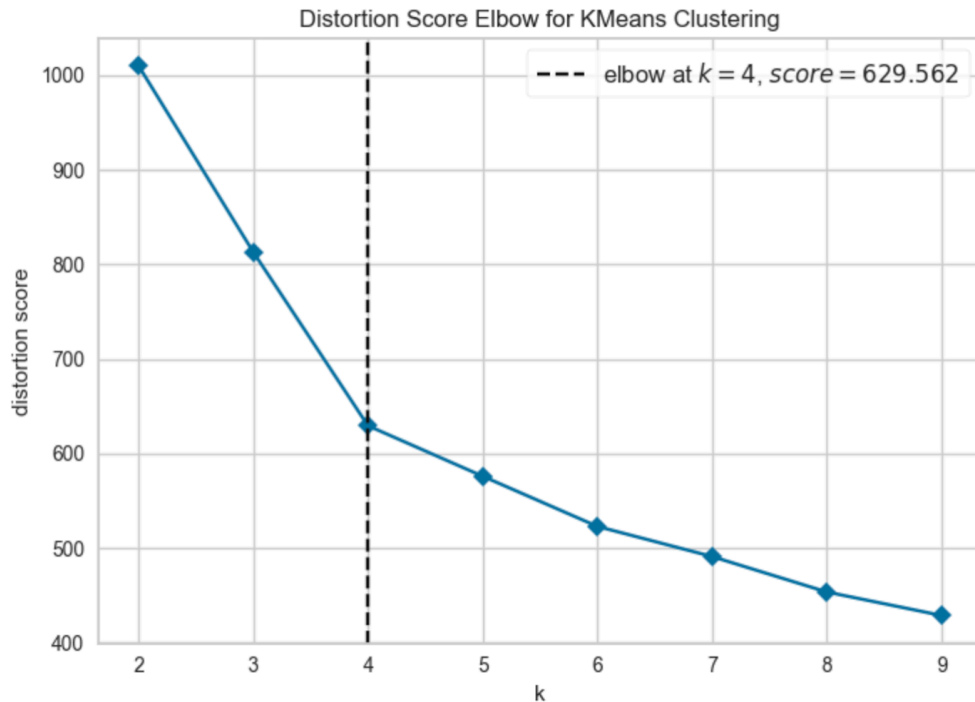
PC4:
-----
Education_Code       0.819312
Age                 0.391927
Family_Size         0.363561
Marital_Status_Code 0.180867
Total_Spending       0.090888
Seniority_Code       0.036878
Name: PC4, dtype: float64

Cumulative variance with 4 components: 89.19%

```

Per triar quines variables utilitzar al clustering hem triat aquelles que tinguin un pes superior a 0.5 en alguna component, això és perquè a partir de 0.5 es considera que la variable té un efecte moderat considerable dins la component. 0-0.2 molt feble, 0.2-0.3 feble, 0.3-0.5 moderat >0.5 molt fort. Observant la taula comparativa de resultats que s'observa en el punt 8 apreciem com aquest llindar és el que ens proporciona millors resultats en quant a mètriques del clustering.

## 6.Determinar k Òptima



### Mètode del colze per trobar la k òptima en KMeans

El mètode del colze consisteix a calcular la inèrcia per diferents valors de  $k$ . La inèrcia és la suma de les distàncies quadrades entre cada punt i el centroid del seu clúster. A mesura que augmenta  $k$ , la inèrcia sempre baixa, però ho fa cada vegada menys.

Per trobar la  $k$  òptima es busca el punt on la reducció de la distorsió comença a ser molt menor. Aquest punt és el "colze". Abans d'aquest punt, afegir clústers millora molt el model; després, afegir-ne més aporta una millora molt petita i només complica el model.

A la gràfica que tenim, la inèrcia baixa molt de  $k=2$  a  $k=3$  i de  $k=3$  a  $k=4$ , però a partir de  $k=4$  la corba s'aplana i les millores són molt petites. Per això el colze es considera a  $k = 4$ , que és el valor on el canvi de pendent és més evident.

Per triar la  $k$  òptima amb l'ull humà és bastant intuïtiu i fàcil, però els ordinadors necessiten una fórmula. Per automatitzar-ho fem servir, Yellowbrick, que fa servir l'algorisme "Kneedle" (Knee Point Detection Algorithm). Aquest mètode dibuixa la corba real (La línia blava que baixa, la inèrcia). Després dibuixa una línia recta on connecta el primer punt ( $K=2$ ) amb l'últim ( $K=11$ ). Per a cada valor de  $K$ , mesura la distància vertical entre la corba real i la línia recta. Finalment, el punt on la corba està més allunyada de la línia recta és el punt de màxima curvatura. Això correspon matemàticament al punt on guanyar més clústers ja no surt a compte. Per tant, no tria a l'atzar. Tria el punt on la millora de qualitat comença a frenar-se en sec.

## 7. Validació del Clustering

Per avaluar la qualitat dels diferents models de clustering, s'han utilitzat diverses mètriques que permeten analitzar tant la cohesió interna dels clústers com la separació entre ells.

### 7.1 Cohesió (SSE)

La cohesió mesura com d'ajustats estan els punts al centroid del seu clúster. S'utilitza la SSE (*Sum of Squared Errors*), que calcula la suma de les distàncies quadrades dels punts al seu centre.

- Valors baixos de SSE indiquen clústers compactes i coherents. Per tant, com més baixa millor.

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

### 7.2 Separació (BSS)

La separació indica fins a quin punt els clústers estan allunyats entre ells. Per mesurar-la s'ha utilitzat una versió robusta del BSS (*Between-Cluster Sum of Squares*), basada en les distàncies entre centres.

- Un valor alt de BSS implica que els clústers estan ben separats i són fàcils de distingir. Per tant, com més alta millor.

$$BSS = \sum_i |C_i| (m - m_i)^2$$

### 7.3 Correlació Incidència-Proximitat

Aquesta mètrica valida la fidelitat estructural dels clústers respecte a la realitat de les dades. Per calcular-la, hem generat dues matrius a partir d'una mostra representativa (1.000 punts):

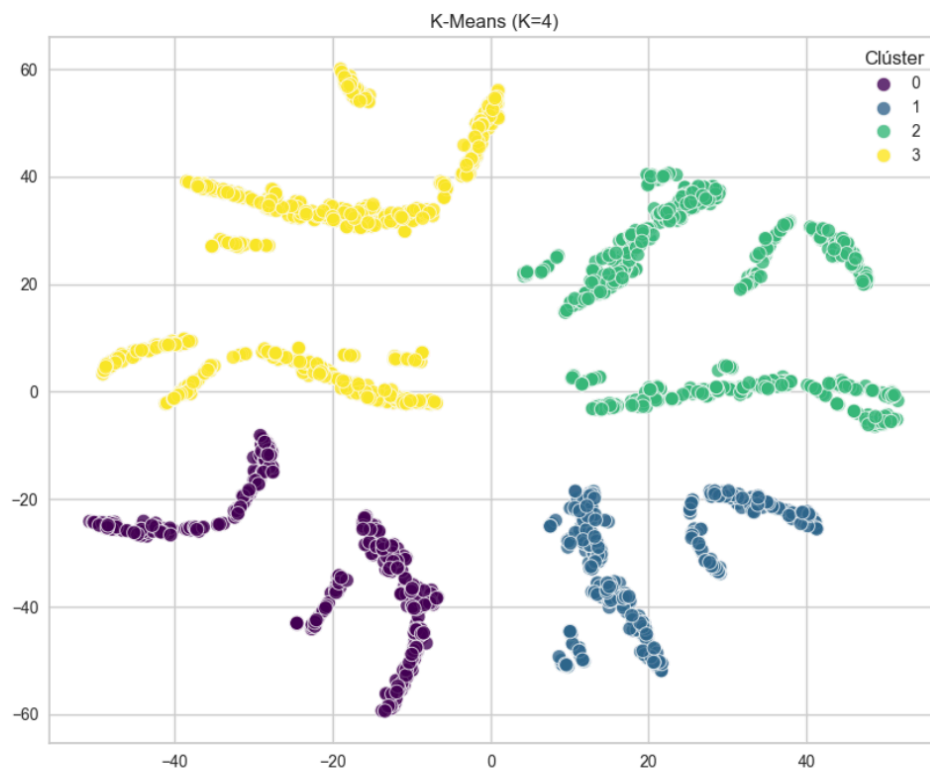
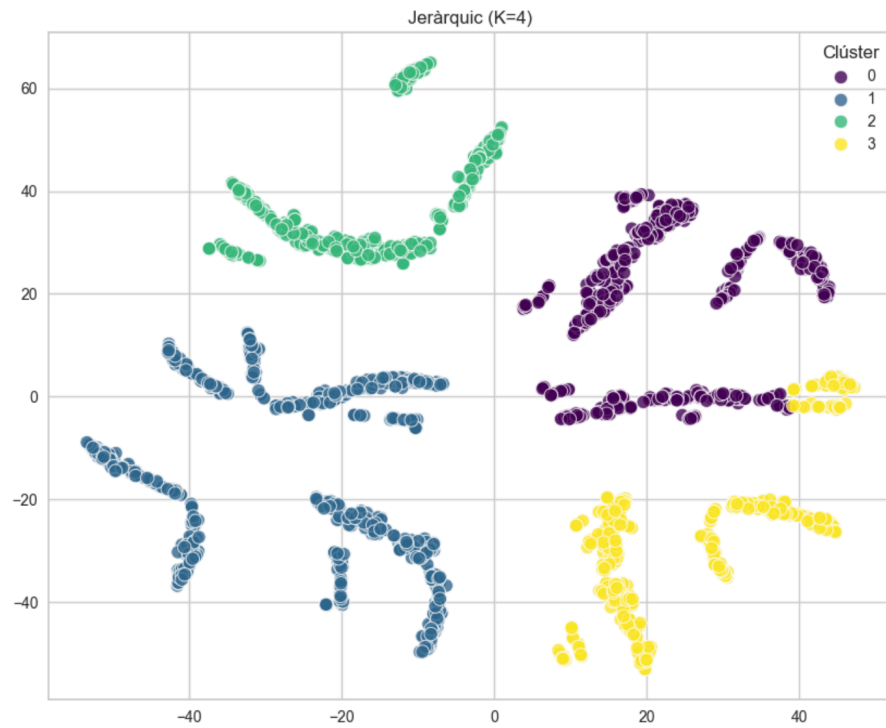
**Matriu d'Incidència (Cij):** Variable binària que val 1 si dos punts pertanyen al mateix clúster i 0 si són en grups diferents.

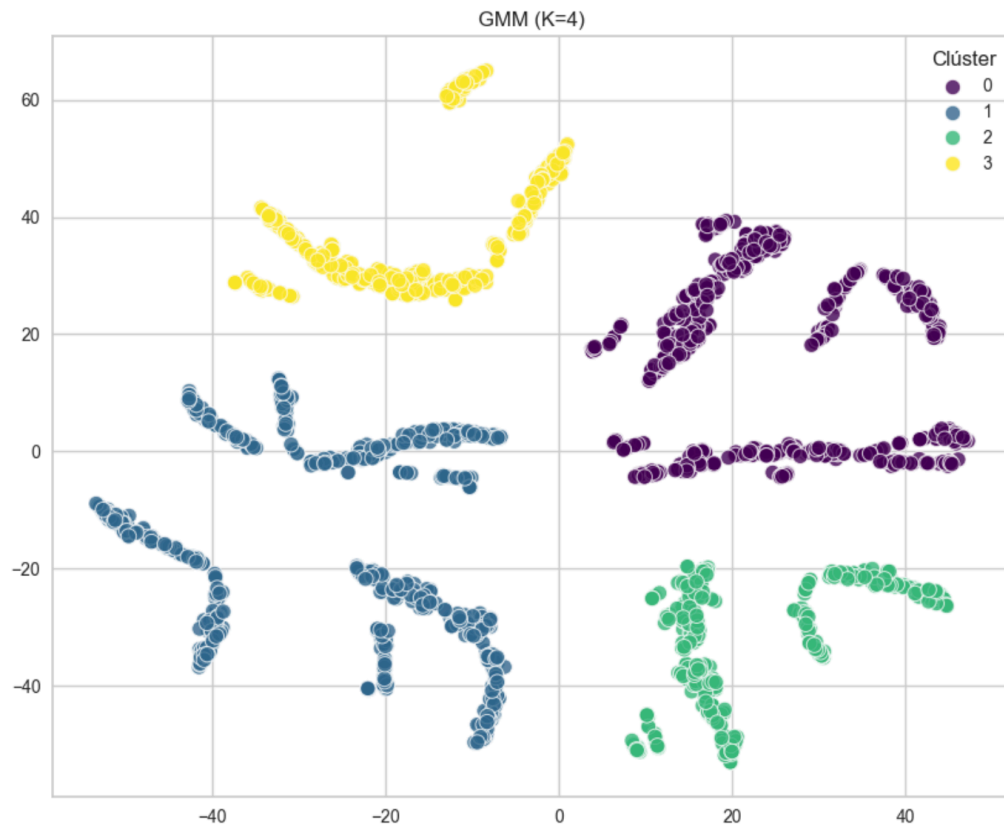
---

**Matriu de Proximitat ( $D_{ij}$ ):** Variable contínua que mesura la distància Euclidiana real entre els dos punts.

Es calcula la Correlació de Pearson entre ambdues matrius. Interpretació: Busquem una relació inversa. L'ideal és que si dos punts són al mateix clúster (1), la seva distància sigui molt petita. Per tant, l'indicador de qualitat òptim és una correlació negativa forta (valors propers a -1). Això confirma que l'algoritme ha respectat la topologia natural de les dades i no ha forçat agrupacions artificials.

## 8. Resultats





Threshold: 0.5	MinMaxScaler			StandardScaler		
Variables	'Income', 'Total_Spending', 'Seniority_Code', 'Education_Code', 'Marital_Status_Code'			'Income', 'Total_Spending', 'Family_Size', 'Seniority_Code', 'Education_Code', 'Marital_Status_Code'		
Algorisme	KMeans	Jeràrquic	GMM	KMeans	Jeràrquic	GMM
SEE	419.93	422.15	416.11	6388.51	6838.00	8889.74
BSS	6.12	6.14	6.23	27.34	27.07	22.00
Correlació	-0.722	-0.727	-0.736	-0.5066	-0.4794	-0.3154

--- Resum de Mètriques ---

K-Means: SSE=419.93, BSS=6.12, Corr=-0.7222

Jeràrquic: SSE=422.15, BSS=6.14, Corr=-0.7278

GMM: SSE=416.11, BSS=6.23, Corr=-0.7365

0,5 i min max

```

--- Resum de Mètriques ---
K-Means: SSE=6388.51, BSS=27.34, Corr=-0.5066
Jeràrquic: SSE=6838.00, BSS=27.07, Corr=-0.4794
GMM: SSE=8889.74, BSS=22.00, Corr=-0.3154

```

0,5 standard

Threshold: 0.3	MinMaxScaler			StandardScaler		
Variables	'Income', 'Total_Spending', 'Family_Size', 'Seniority_Code', 'Education_Code', 'Marital_Status_Code', 'Age'			'Income', 'Total_Spending', 'Family_Size', 'Seniority_Code', 'Education_Code', 'Marital_Status_Code', 'Age'		
Algorisme	KMeans	Jeràrquic	GMM	KMeans	Jeràrquic	GMM
SEE	629.58	633.17	633.17	8366.79	9324.57	9950.69
BSS	6.34	6.40	6.40	27.84	25.65	24.77
Correlació	-0.703	-0.713	-0.714	-0.4710	-0.3829	-0.3936

```

--- Resum de Mètriques ---
K-Means: SSE=629.58, BSS=6.34, Corr=-0.7030
Jeràrquic: SSE=633.17, BSS=6.40, Corr=-0.7130
GMM: SSE=633.17, BSS=6.40, Corr=-0.7141

```

0,3 minmax (inclou age i family\_size)

```

--- Resum de Mètriques ---
K-Means: SSE=8366.79, BSS=27.84, Corr=-0.4710
Jeràrquic: SSE=9324.57, BSS=25.65, Corr=-0.3829
GMM: SSE=9950.69, BSS=24.77, Corr=-0.3936

```

0,3 standard (inclou age)

## 9. Clústers

Mirem amb el millor model (minmax) i 0.5.

### 9.1 Clústers amb KMeans

PERFIL COMPLET: K-MEANS				
KMeans_Cluster	0	1	2	3
Income	51091.59	47608.09	52155.99	56888.53
Age	56.80	55.70	55.03	56.39
Family_Size	2.32	3.04	2.89	2.03
Seniority_Code	2.00	1.00	2.00	1.00
Education_Code	3.38	3.44	3.45	3.56
Marital_Status_Code	2.59	1.29	1.00	2.78
Total_Spending	689.46	394.10	715.23	683.83
Recency	50.12	48.43	49.23	47.99
MntWines	350.68	196.57	370.50	334.23
MntFruits	28.91	17.95	29.94	31.35
MntMeatProducts	184.70	105.13	188.78	198.46
MntFishProducts	43.45	25.35	41.47	43.93
MntSweetProducts	29.93	18.71	31.30	31.49
MntGoldProds	51.79	30.40	53.24	44.38
NumWebPurchases	4.59	3.38	4.62	3.98
NumCatalogPurchases	2.89	1.88	2.97	3.09
NumStorePurchases	6.16	5.17	6.18	6.03
NumWebVisitsMonth	5.91	5.12	5.90	4.37
Count (Clients)	649.00	617.00	398.00	500.00

#### Cluster 0

Family\_Size petit (2.3).

Despesa total alta (690) amb força vi, carn i productes Gold.

Bastantes compres web, catàleg i botiga.

Llars petites amb alt nivell de despesa i ús intensiu de tots els canals.

- Programa de Subscripció Premium (Tipus "Prime"): Com que compren molt sovint i per tots els canals, ofereix-los una subscripció anual amb enviaments gratuïts i accés anticipat a productes exclusius (especialment vi i productes Gold). Això assegura la seva fidelitat i augmenta el tiquet mitjà.
- Venda Creuada de Luxe (Cross-selling): Ja que compren carn i vi de qualitat, envia'ls recomanacions personalitzades de maridatge (ex: "Has comprat aquest entrecot, prova aquest vi reserva amb un 10% de descompte").

### Cluster 1

Family\_Size més gran (3.0).

Despesa total més baixa de tots (394) i menys despesa en totes les categories.

Menys compres i visites web.

Famílies més grans amb comportament de compra modest, segment de baix valor.

- Packs Familiars i Format Estalvi: Crea lots de productes bàsics (carn, peix, fruita) en formats grans (ex: "Pack familiar 5kg") amb un preu per quilo molt competitiu. Això els ajuda a omplir el rebost sense disparar la despesa.
- Cupons de "Tornada a la Botiga": Com que tenen poca activitat, envia'ls cupons de descompte agressius (ex: "10€ de regal en la teva propera compra superior a 50€") que caduquin ràpidament per reactivar-los.

### Cluster 2

Family\_Size mitjà (2.9).

Despesa total més alta (715), màxim consum en vi, carn i Gold.

És dels que més compren per web, catàleg i botiga.

Segment de clients top, molt rendibles i molt actius en tots els canals.

- Club d'Excel·lència / Esdeveniments Exclusius: No els bombardegis amb descomptes (ja compren igualment). Convida'ls a tastos de vins exclusius o presentacions de productes Gold a la botiga física. Fes-los sentir especials per mantenir la connexió emocional amb la marca.
- Gamificació a l'App/Web: Com que són molt digitals, crea un sistema de punts on, per cada compra web o ressenya, acumulin saldo per a productes de gamma alta. Incentiva que segueixin comprant per "pujar de nivell".

### Cluster 3

Family\_Size més petit (2.0).

Despesa total també alta (684), similar a 0.

Compres i visites web una mica per sota de 0 i 2.

Llars petites amb alta despesa, però una mica menys digitals que el clúster 2.

- Digitalització Incentivada: El seu punt feble és la web. Ofereix-los un descompte exclusiu només si fan la seva primera compra online o si descarreguen l'App ("15% de descompte en la teva primera comanda web"). L'objectiu és moure'ls cap al canal digital on és més fàcil vendre'ls més coses.
- Catàleg Físic amb QR: Si responen millor a canals tradicionals, envia'ls un catàleg a casa molt visual de vins i productes gourmet, però amb codis QR que els portin directament a la fitxa del producte per facilitar la compra ràpida.

## 9.2 Clústers amb Jeràrquic

PERFIL COMPLET: JERÀRQUIC				
Hierarchical_Cluster	0	1	2	3
Income	51641.00	51600.31	52066.57	51319.82
Age	56.50	56.05	55.94	55.68
Family_Size	2.23	2.95	1.90	2.92
Seniority_Code	2.00	1.00	1.00	2.00
Education_Code	3.24	3.48	3.53	3.61
Marital_Status_Code	2.67	1.40	3.00	1.16
Total_Spending	718.66	519.34	532.16	675.61
Recency	50.59	48.63	47.49	48.81
MntWines	359.59	257.47	259.53	356.54
MntFruits	31.15	23.26	25.24	27.04
MntMeatProducts	195.63	145.78	149.02	174.83
MntFishProducts	46.39	33.14	34.66	38.21
MntSweetProducts	31.75	23.78	25.65	28.87
MntGoldProds	54.15	35.91	38.05	50.13
NumWebPurchases	4.55	3.69	3.57	4.66
NumCatalogPurchases	2.98	2.42	2.42	2.84
NumStorePurchases	6.22	5.67	5.34	6.10
NumWebVisitsMonth	5.83	4.80	4.77	6.01
Count (Clients)	575.00	729.00	388.00	472.00

Aquí Age i Income són pràcticament iguals a tots els grups; el que canvia és bàsicament la despesa i una mica la mida de la llar.

### Cluster 0

Family\_Size (2.2)

Despesa total més alta (719) i molta despesa en totes les categories.

Moltes compres i visites web.

Clients molt de valor, molt actius.

### Cluster 1

Family\_Size (3)

Despesa total mitjana (519), per sota de 0 i 3.

Patró de compres moderat.

Famílies més grans amb despesa mitjana.

### Cluster 2

Family\_Size més petit (1.9).

Despesa total mitjana (532), semblant a 1.

Patró de compra també moderat.

Llars petites amb despesa mitjana.

### Cluster 3

Family\_Size (2.9)

Despesa total alta (676), bastant vi i productes Gold.

És qui té més visites web.

Clients de valor alt i molt actius digitalment.

### 9.3 Clústers amb GMM

=====				
PERFIL COMPLET: GMM				
=====				
GMM_Cluster	0	1	2	3
Income	51091.59	51600.31	52155.99	52066.57
Age	56.80	56.05	55.03	55.94
Family_Size	2.32	2.95	2.89	1.90
Seniority_Code	2.00	1.00	2.00	1.00
Education_Code	3.38	3.48	3.45	3.53
Marital_Status_Code	2.59	1.40	1.00	3.00
Total_Spending	689.46	519.34	715.23	532.16
Recency	50.12	48.63	49.23	47.49
MntWines	350.68	257.47	370.50	259.53
MntFruits	28.91	23.26	29.94	25.24
MntMeatProducts	184.70	145.78	188.78	149.02
MntFishProducts	43.45	33.14	41.47	34.66
MntSweetProducts	29.93	23.78	31.30	25.65
MntGoldProds	51.79	35.91	53.24	38.05
NumWebPurchases	4.59	3.69	4.62	3.57
NumCatalogPurchases	2.89	2.42	2.97	2.42
NumStorePurchases	6.16	5.67	6.18	5.34
NumWebVisitsMonth	5.91	4.80	5.90	4.77
Count (Clients)	649.00	729.00	398.00	388.00

#### Cluster 0

Igual que el K-means 0: Family\_Size petit, alta despesa (689), molta activitat en tots els canals.

Llars petites de alt valor.

#### Cluster 1

Molt semblant al Jeràrquic 1: Family\_Size gran (2.95), despesa mitjana (519), patró de compra moderat.

Famílies més grans de valor mitjà/baix.

#### Cluster 2

Igual que el K-means 2: màxima despesa (715) i molt consum en vi, carn i Gold; molta activitat en tots els canals.

Segment top clar, clients més rendibles.

#### Cluster 3

Pràcticament igual al Jeràrquic 2: Family\_Size més petit (1.9), despesa mitjana (532), patró de compra moderat.

Llars petites amb despesa mitjana.

