

Aquest dataset conté 2.240 files (clients) i 29 columnes, amb una combinació de dades numèriques, categòriques i temporals. És ric però brut, amb moltes inconsistències i soroll, per la qual cosa la prioritat és la neteja.

#### 1. Estructura general:

- **Dades numèriques:** Ingressos, despeses en productes, nombre de visites web, any de naixement, etc.
- **Dades categòriques:** Nivell educatiu i estat civil.
- **Dades temporals:** Data d'alta del client (es convertirà en antiguitat en dies si es necessita en el futur).

#### 2. Anàlisi de les variables categòriques:

- **Nivell educatiu:** Està descompensat amb un gran nombre d'educats universitaris. El grup **Basic** és molt petit, per la qual cosa podríem agrupar-lo amb **2n Cycle** i **Master** per simplificar-lo.
- **Estat civil:** La columna **Marital\_Status** té valors inusuals com "**Alone**", "**Absurd**", i "**YOLO**" que s'han eliminat, ja que no són vàlids.

#### 3. Anàlisi de les variables numèriques:

- **Ingressos:** És la variable més rellevant per segmentar els clients segons el seu poder adquisitiu.
- **Despesa:** Algunes persones tenen **valors de despesa zero** en productes, la qual cosa indica que no han comprat aquests productes.
- **Famílies:** Les variables **Kidhome** i **Teenhome** són útils per identificar si el client té fills o adolescents.

#### 4. Qualitat de les dades: Nulls i Outliers:

- **Valors nuls:** Hi ha **24 files amb ingressos nuls**, que es decideix eliminar perquè representen només un **1% de les dades**.
- **Outliers:**
  - **Edat:** Alguns clients tenen **edats impossibles** (més de 120 anys). S'elimina qualsevol client nascut abans de 1920.
  - **Ingressos extrems:** Es detecta un client amb **666.666 d'ingressos**, que es considera un outlier i es elimina per evitar distorsionar el

clustering.

Aquest procés de neteja prepara el dataset per aplicar tècniques de clustering de manera més fiable, evitant errors i assegurant-se que les dades siguin representatives per als models.

## Products

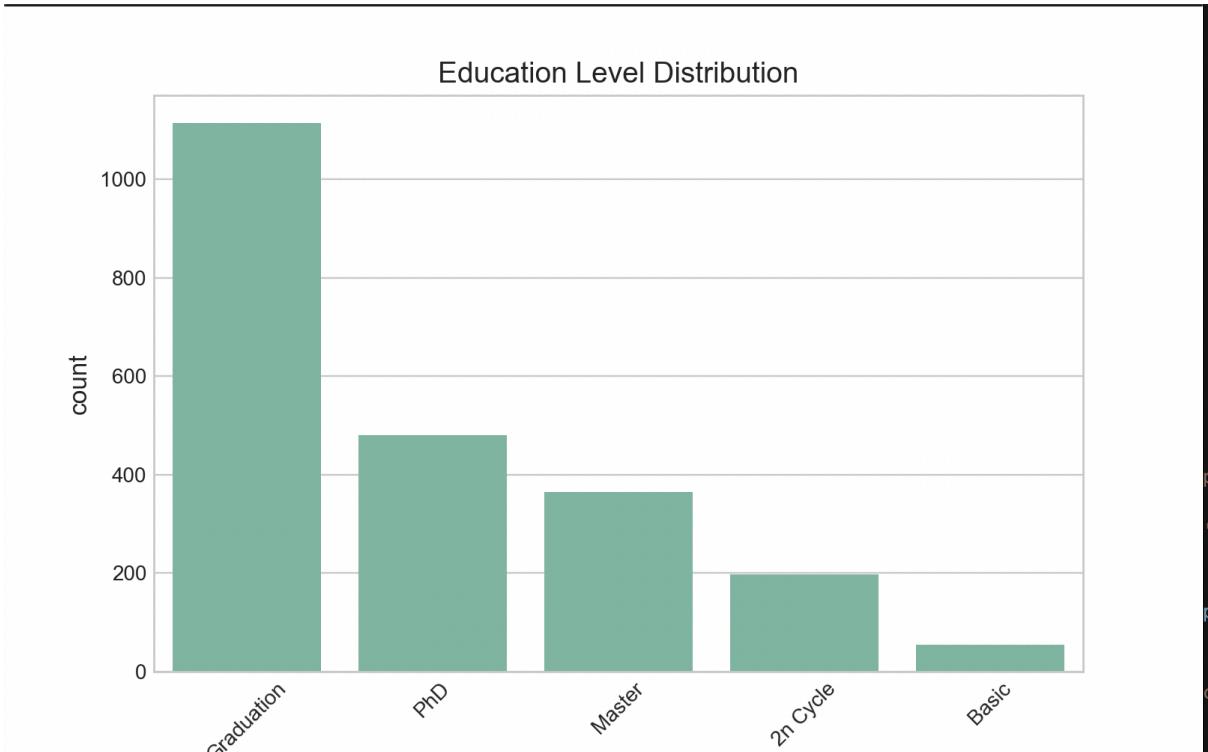
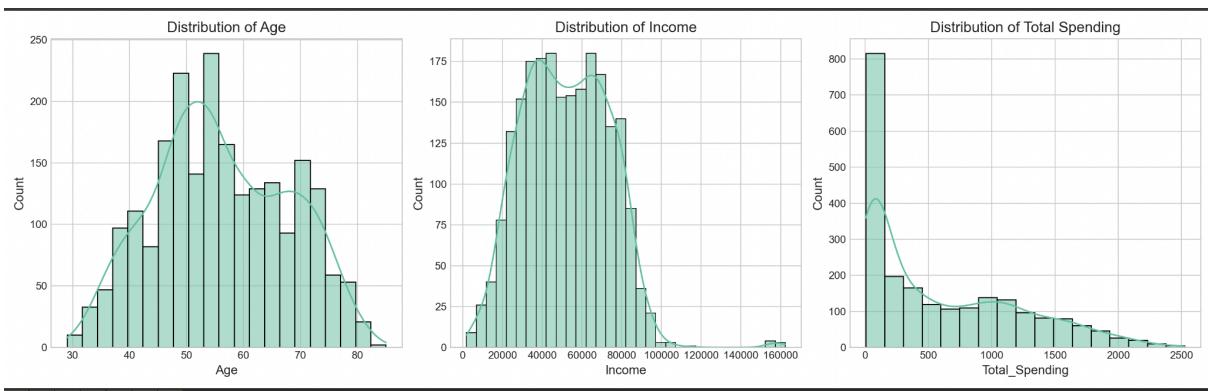
- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

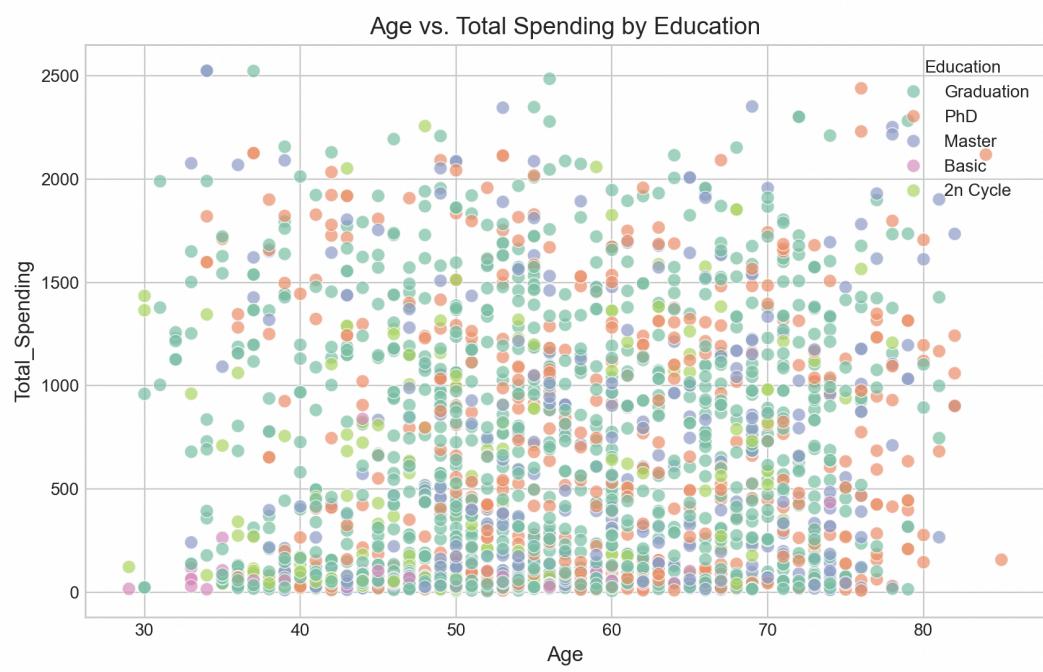
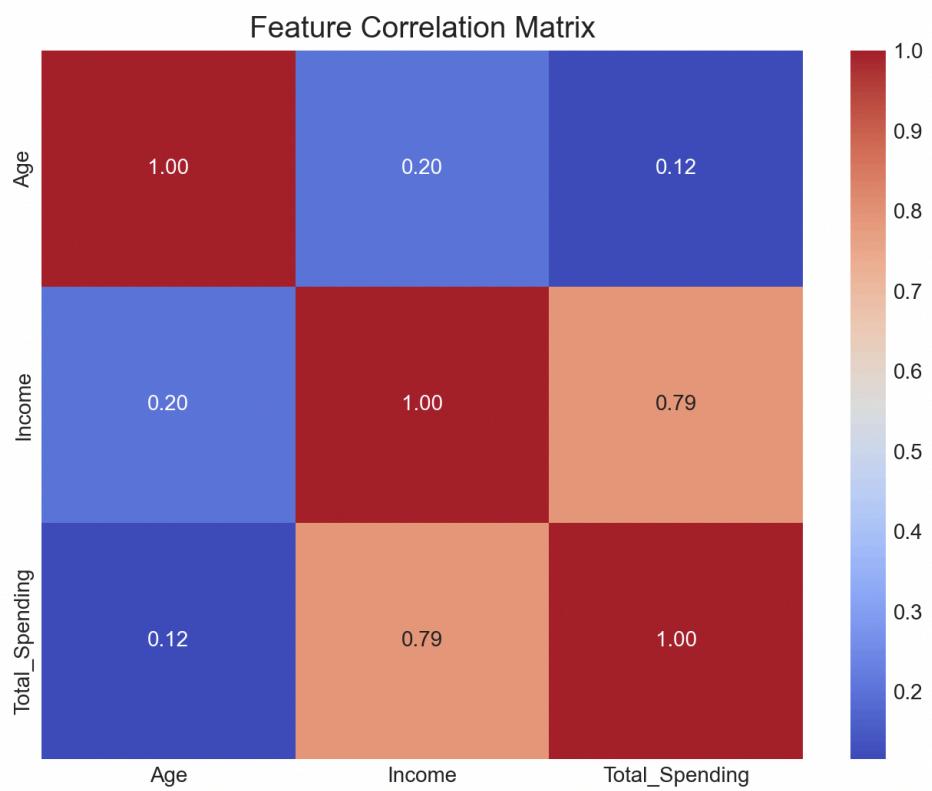
## Promotion

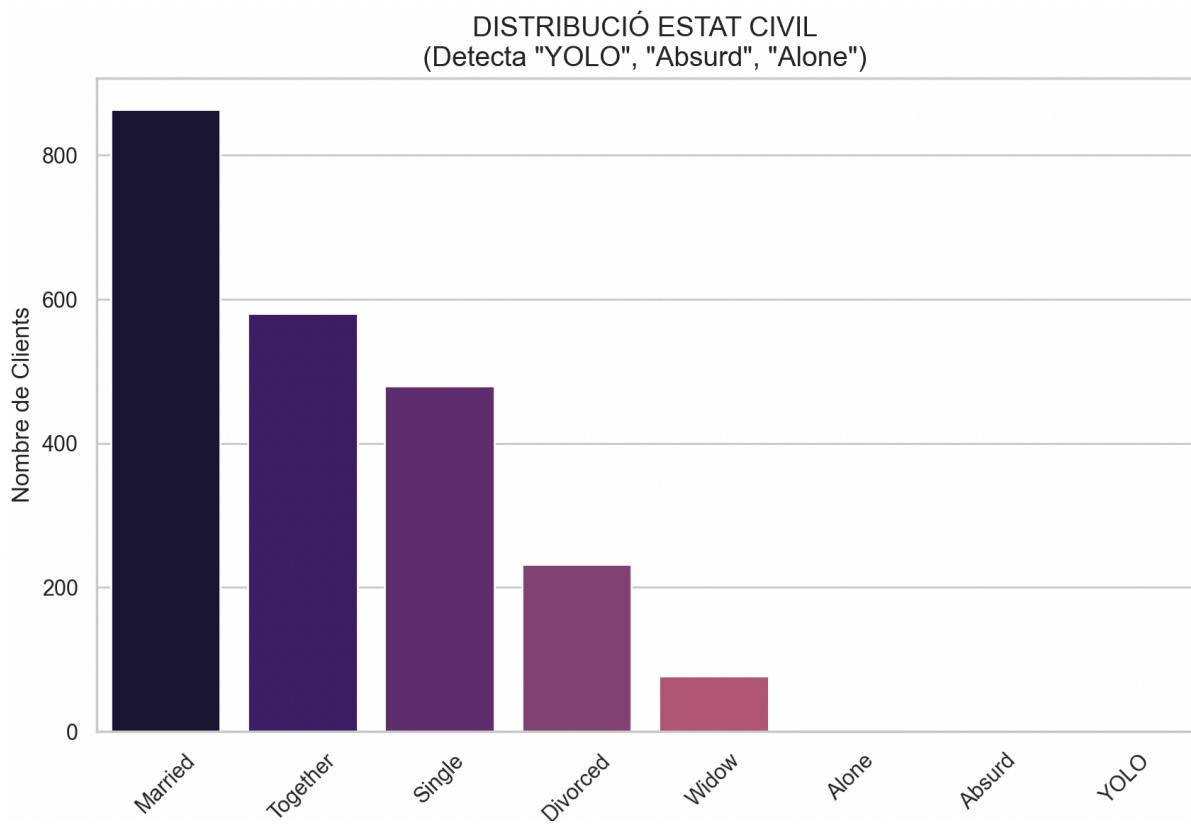
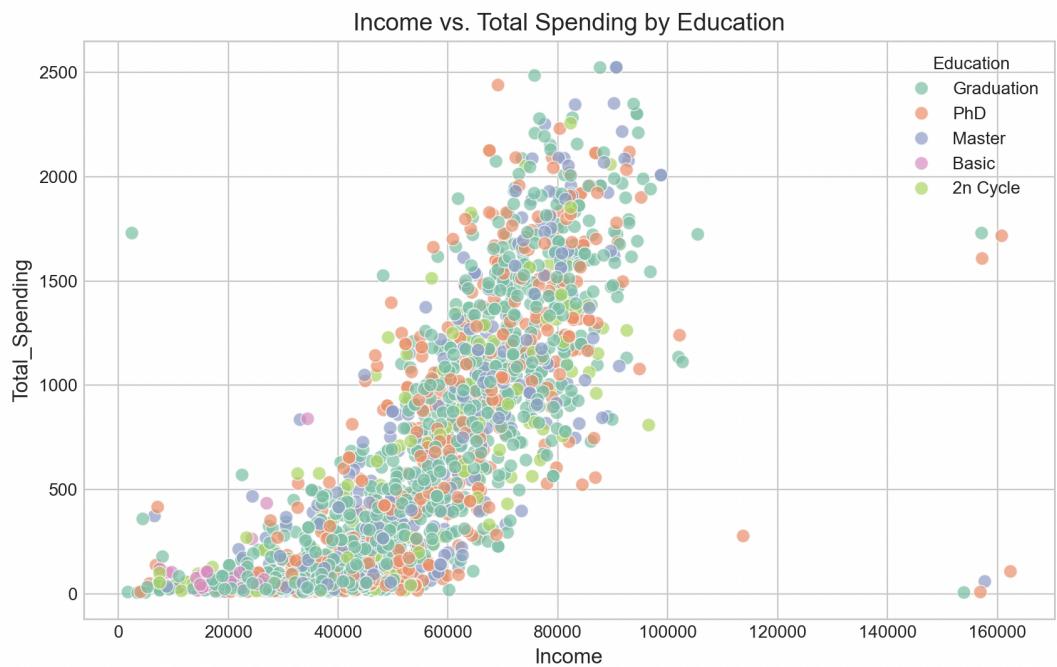
- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

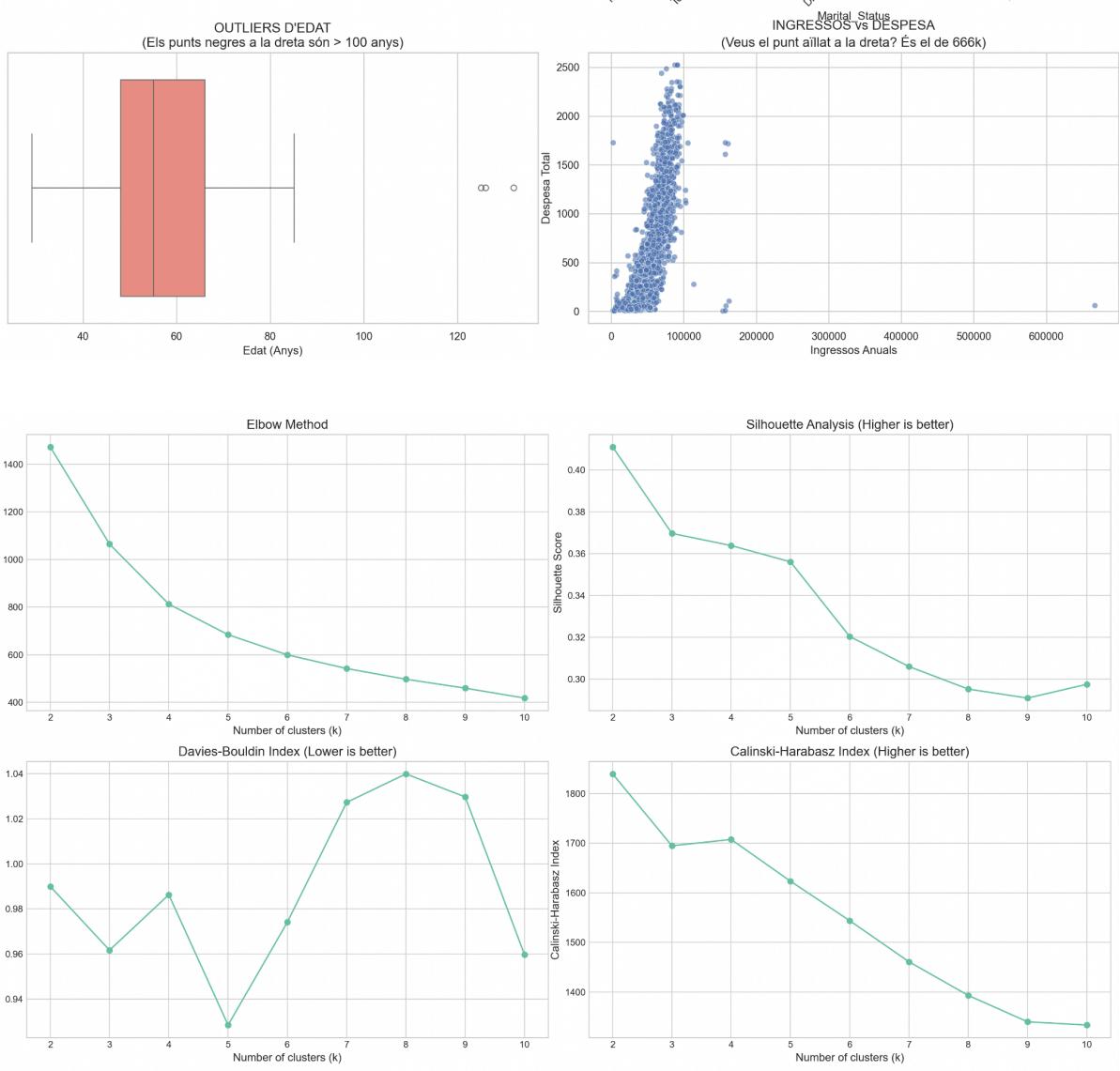
## Place

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month









## Validar clustering:

Cohesió (SSE): Calcula quant de "compactes" són els grups. Com més baix, millor (els punts estan molt a prop del seu centre).

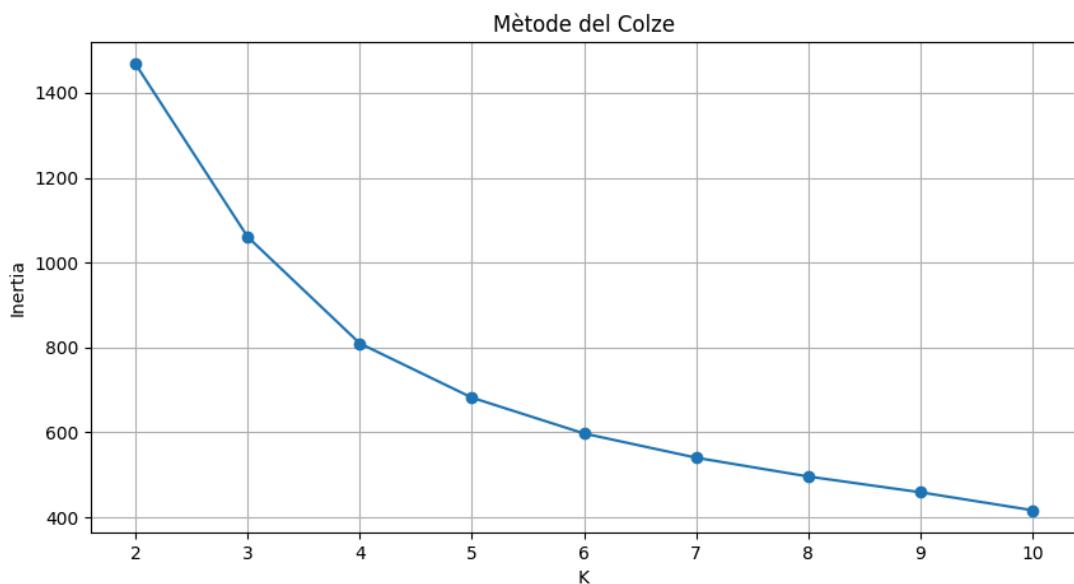
Separació (BSS): He implementat una versió robusta basada en la distància entre els centres dels clústers. Com més alt, millor (els grups estan ben separats entre ells).

Correlació: Fa servir una mostra de 1.000 punts (perquè no trigui una eternitat) i mira si els punts que l'algorisme ha posat junts realment estan a prop en l'espai original. Un valor proper a -1 és perfecte (vol dir que incidència 1 es correspon amb distància petita).

Mapa de Calor (Heatmap): Reordena els clients segons el seu grup i pinta la distància. Si el clustering és bo, hauries de veure quadrats foscos (baixa distància) al llarg de la línia diagonal.

## CONCLUSIÓ INICIAL

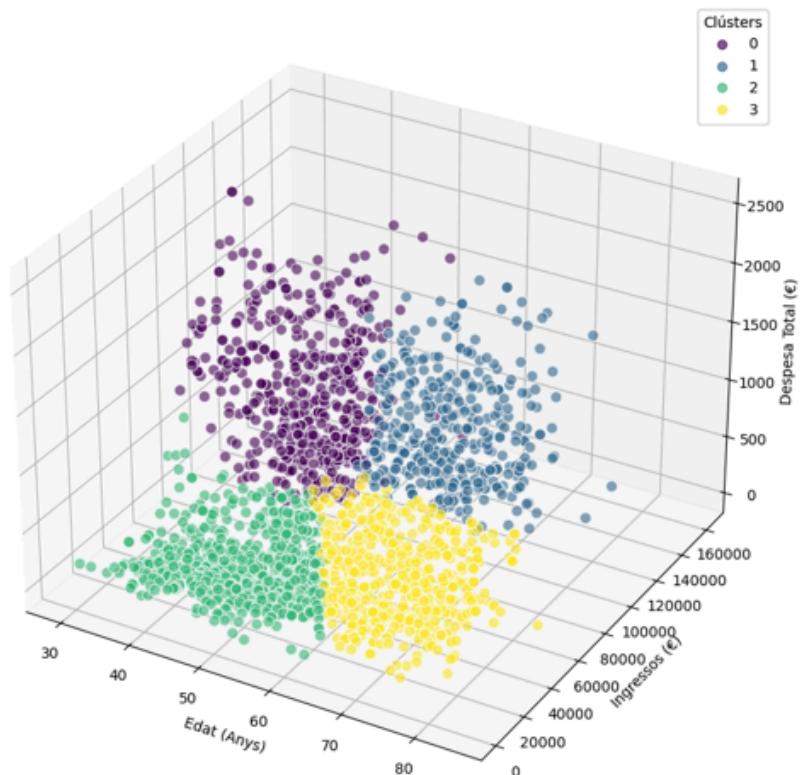
Amb els valors obtinguts de cada mètode podem apreciar com k-means



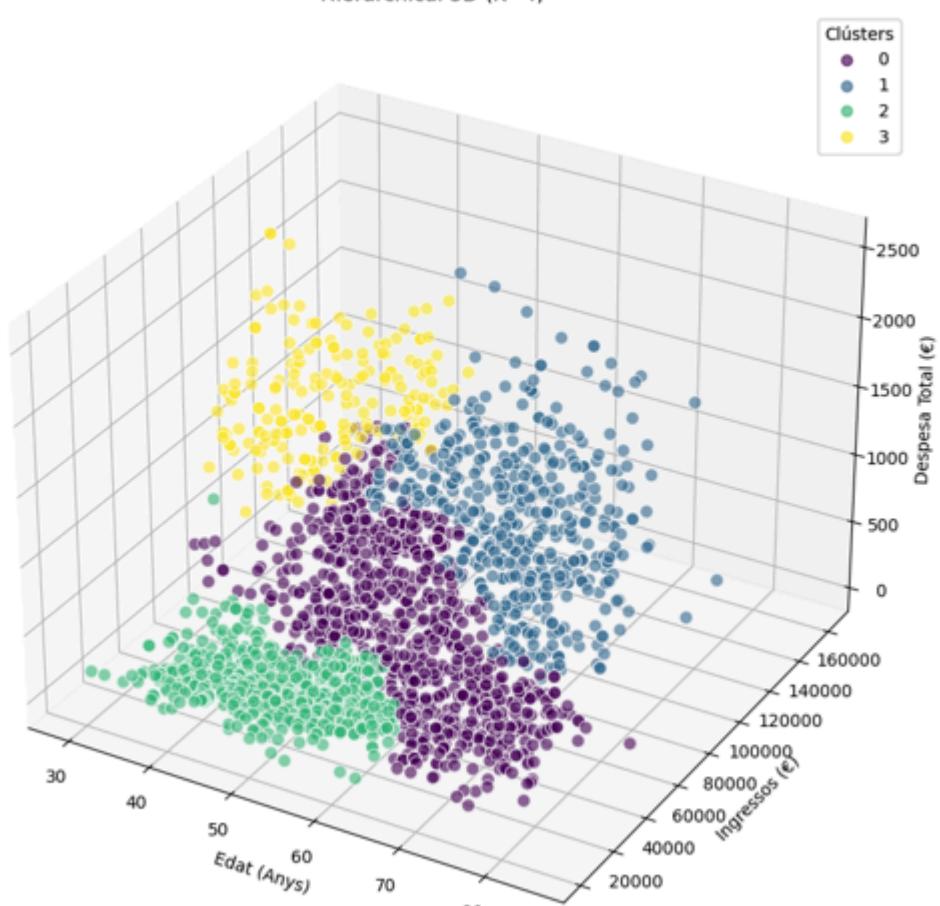
```
--- Preprocessing & Cleaning ---  
--- Cerca de K Òptima ---  
La K òptima seleccionada és: 4  
--- Executant i Guardant Models (K=4) ---  
Processant K-Means...  
-> Gràfic 3D guardat: 01_kmeans_3d.png  
Processant Jeràrquic...  
-> Gràfic 3D guardat: 02_hierarchical_3d.png  
Processant GMM...  
-> Gràfic 3D guardat: 03_gmm_3d.png  
--- Resum de Mètriques ---  
K-Means: SSE(Cohesió)=809.74, BSS(Separació)=8.88, Correlació=-0.5681  
Jeràrquic: SSE(Cohesió)=1032.59, BSS(Separació)=9.28, Correlació=-0.5131  
GMM: SSE(Cohesió)=1404.60, BSS(Separació)=7.34, Correlació=-0.4190  
CSV guardat a: resultats_clustering/marketing_campaign_final.csv  
Processament completat.
```

Amb els valors obtinguts de cada mètode podem apreciar com k-means té un SSE menor indicant que és el mètode que genera uns clústers amb major cohesió, la distància entre els punts d'un mateix clúster és més baixa. En canvi veiem com el mètode jeràrquic té una millor separació entre clústers, però no gaire més gran que k-means, per últim La correlació entre la matriu d'incidència i la matriu de proximitat és negativa tant en k-means com en el mètode jeràrquic, la qual cosa suggereix que hi ha una certa desconformitat entre els clusters i la proximitat dels punts dins de la matriu. K-Means sembla ser el millor algorisme per a aquest conjunt de dades, ja que mostra una bona cohesió i una separació raonable. Tot i la correlació negativa, aquest mètode és el més adequat per a la segmentació de clients. Gaussian Mixture Models no és adequat per a aquest conjunt de dades, ja que presenta la cohesió més feble, la separació més baixa i la correlació més negativa.

K-Means 3D (K=4)



Hierarchical 3D (K=4)



GMM 3D (K=4)

