1. Prétraitement et ingénierie RFM

3. Analyse de la distribution : Observez l'asymétrie (skewness) de vos variables. Pourquoi une simple standardisation suffit-elle rarement sur des données de montant ?

1. Distribution fortement asymétrique (skewness = 19.32)
2. Présence de valeurs extrêmes (outliers)
3. StandardScaler assume une distribution normale
4. Les outliers dominent la variance et biaisent l'analyse
5. SOLUTION : Transformation logarithmique + standardisation

2. Réduction de dimension par PCA

Pourquoi une variable avec une variance immense (ex: le montant total en euros) écraserait-elle les autres variables (ex: le nombre de commandes) si on ne standardise pas les données avant la PCA ?

La PCA cherche des axes qui maximisent la **variance**. Si une variable (ex: **montant en €**) varie de 0 à 50 000 alors qu'une autre (ex: **nombre de commandes**) varie de 1 à 20, la matrice de covariance est dominée par la grosse échelle :

- le premier axe (PC1) va s'aligner presque entièrement sur la variable « montant »,
- les autres variables deviennent du « bruit » dans la décomposition.

Sans standardisation, la PCA fait surtout une **PCA des unités de mesure**, pas des comportements.

Si le déterminant de votre matrice de covariance est proche de zéro, qu'est-ce que cela indique sur la relation entre vos variables originales ?

Un déterminant proche de 0 signifie que la matrice est **quasi singulière** :

- Il existe une ou plusieurs **relations linéaires fortes** entre variables (colinéarité),
- Donc une ou plusieurs dimensions n'apportent presque pas d'information indépendante.

Traduction métier : les variables racontent en partie la même histoire (ex: M très corrélé à F si « plus on achète souvent, plus on dépense »).
Traduction PCA : certaines composantes auront une **valeur propre d'environ 0** → dimension « inutile » à garder.

5. Interprétation et application métier

Analyse qualitative : Prenez 5 produits au hasard dans un même cluster. Leurs descriptions semblent-elles traiter du même univers ?

5 produits dans un cluster :

- HOME SWEET HOME BOTTLE
- STRAWBERRY HONEYCOMB GARLAND
- CHARLIE AND LOLA TABLE TINS
- METAL SIGN EMPIRE TEA
- TROPICAL HONEYCOMB PAPER GARLAND

Mots-clés : art, wall art, wall, metal sign, sign, metal, home, flower, hot, water bottle

5 produits dans un autre cluster :

- LUSH GREENS RIBBONS
- DAISY FOLKART HEART DECORATION
- BLACK/BLUE POLKADOT UMBRELLA
- HYACINTH BULB T-LIGHT CANDLES
- check

Mots-clés : set, check, pink, blue, heart, vintage, red, christmas, bag, glass

Esprit critique : Pourquoi le Deep Learning est-il ici plus puissant qu'un simple clustering sur les mots ?

1. Apprentissage de représentations latentes abstraites
2. Capture relations non-linéaires entre mots
3. Comprend sémantique : 'red rose' ≈ 'pink flower'
4. Gère automatiquement synonymes et variations
5. Robuste aux erreurs orthographiques
6. Généralise mieux à nouveaux produits

Quelles sont les limites de cette approche si les descriptions sont trop courtes (ex: "Blue Vase") ?

Problèmes :

- Contexte insuffisant (2 mots)
- Ambiguïté fonctionnelle
- Peu d'information pour apprentissage

Solutions :

1. Enrichir avec métadonnées (catégorie, prix, matériau)
2. Ajouter features visuelles (CNN sur images)
3. Utiliser embeddings pré-entraînés (BERT, Word2Vec)
4. Combiner texte + données numériques