Quel méthode d'échantillonnage utiliser ?

En fonction des exigences détaillées de votre projet, je recommande d'utiliser **l'échantillonnage aléatoire stratifié** pour les raisons suivantes :

1. Vous avez trois catégories de produits distinctes (pantalons, robes, pulls) qui doivent être analysées séparément et comparées.
2. Vos objectifs incluent la compréhension des performances par catégorie et la prise de décisions stratégiques.
3. Vous prévoyez de réaliser des tests d'hypothèses entre les catégories ultérieurement.
4. Les données serviront à des prédictions de ventes, donc préserver la représentativité est crucial.

Taille de l'échantillon : tout ou un pourcentage ?

Étant donné que vos ensembles de données sont très petits (37, 38, 37 entrées), je recommande fortement d'utiliser **l'ensemble des données** plutôt qu'un échantillon. Voici pourquoi :

1. **Considérations pour petits ensembles de données** :
    - Avec seulement ~37 entrées par catégorie, un échantillon réduit diminuerait significativement la puissance statistique.
    - Vous risqueriez de perdre des motifs ou des variations importants dans les données.
    - Pour l'apprentissage automatique et l'analyse statistique (que vous envisagez), avoir plus de points de données fournira des résultats plus fiables.

Expliquer le Z-score dans ce contexte

Le Z-score sert à détecter les valeurs aberrantes dans l'analyse de vos données de ventes.
Il mesure à combien d'écarts-types un point de données se situe par rapport à la moyenne, c'est-à-dire à quel point une valeur est "inhabituelle" dans votre jeu de données.

La formule du Z-score est :

$$\mathbf{Z = (X - \mu) / \sigma}$$

Où :

- $\mathbf{X}$ = valeur analysée
- $\boldsymbol{\mu}$ = moyenne de l'ensemble de données
- $\boldsymbol{\sigma}$ = écart-type de l'ensemble de données

---

Comment choisir le seuil (seuil) ?

Le seuil du Z-score est généralement choisi en fonction de plusieurs critères :

1. **Théorie statistique** :
    - **Z > 3** : Capture environ 0,3 % des valeurs les plus extrêmes (règle des 99,7 %).
    - **Z > 2** : Capture environ 5 % des valeurs les plus extrêmes (règle des 95 %).
    - **Z > 1** : Capture environ 32 % des valeurs les plus extrêmes (règle des 68 %).
2. **Contexte métier** : Pour vos données e-commerce :
    - **Ventes (S)** : Un seuil élevé (3-3,5) pourrait être pertinent, car les ventes varient souvent naturellement.
    - **Pages vues (X4)** : Un seuil standard (3) peut suffire, car les patterns de trafic sont généralement prévisibles.
    - **Indices clients/transactions (X9, X10)** : Un seuil plus bas (2,5) peut être utilisé si une certaine stabilité est attendue.
3. **Taille de l'échantillon** :
    - **Petits jeux de données (<100, comme le vôtre)** : Envisagez des seuils plus bas (2,5-2,7).
    - **Grands jeux de données (>1000)** : Utilisez des seuils standards plus élevés (3-3,5).
    - **Très grands jeux de données (>10000)** : Vous pouvez envisager des seuils encore plus élevés (4+).

Mise à l’échelle des caractéristiques : pourquoi ?

- Les variables de votre dataset (par exemple, ventes, pages vues, indices de transaction) sont exprimées dans des unités différentes : milliards de Yuans, millions, centaines de milliers, etc.
- Ces différences d'échelle peuvent **biaiser les résultats des analyses** (corrélations, ACP, etc.) et **affecter la performance des algorithmes d'apprentissage automatique** (régressions, réseaux neuronaux, etc.).
- Exemple : une variable mesurée en milliards aura une influence disproportionnée par rapport à une autre mesurée en centaines de milliers.