Groupe 1 : Secteur BANCAIRE - Credit Card Dataset

- **Source** : Kaggle - "Credit Card Dataset for Clustering"
- **Taille** : 8 950 clients
- **Features** : 17 variables numériques continues
- **Complexité moyenne** :
    - Toutes variables numériques (pas d'encoding complexe)
    - Quelques outliers à gérer simplement
    - Scaling standard suffisant
    - Interprétation claire des comportements d'usage
- **Lien** : kaggle.com/arjunbhasin2013/ccdata

Groupe 2 : Secteur MARKETING/E-COMMERCE - Marketing Campaign Dataset

- **Source** : Kaggle - "Marketing Campaign"
- **Taille** : 2 240 clients
- **Features** : 29 variables (mix simple : âge, revenus, achats par catégorie, réponse campagnes)
- **Complexité moyenne** :
    - Taille raisonnable pour calculs rapides
    - Variables majoritairement numériques
    - Peu de catégorielles (Education, Marital_Status)
    - Feature engineering simple (age depuis Year_Birth)
- **Lien** : kaggle.com/rodsaldanha/arketing-campaign

Groupe 3 : Secteur AUTOMOBILE - Car Insurance Claim

- **Source** : Kaggle - "Car Insurance Claim Prediction"
- **Taille** : 10 000 clients
- **Features** : 19 variables (données assurés : âge, véhicule, historique sinistres)
- **Complexité moyenne** :
    - Volume modéré
    - Variables catégorielles limitées et simples (Gender, Car_Category)
    - Peu de valeurs manquantes
    - Interprétation métier directe
- **Lien** : kaggle.com/datasets (chercher "car insurance claim")

Groupe 4 : Secteur SANTÉ - Healthcare Provider Fraud Detection

- **Source** : Kaggle - "Healthcare Provider Fraud Detection Analysis"
- **Taille** : ~5 410 prestataires (après agrégation)
- **Features** : 15-20 variables (montants réclamés, diagnostics, procédures, démographie patients)
- **Complexité moyenne** :
    - Agrégation par prestataire simple

- Variables principalement numériques (montants, counts)
- Quelques catégorielles basiques
- PCA efficace sur coûts/volumes

- **Lien** : kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

Groupe 5 : Secteur TOURISME/HÔTELLERIE - Hotel Reservations Dataset

- **Source** : Kaggle - "Hotel Reservations Classification"
- **Taille** : 36 275 réservations
- **Features** : 19 variables (type chambre, lead time, invités, tarifs, statut annulation)
- **Complexité moyenne** :
    - Volume gérable
    - Variables bien structurées
    - Catégorielles limitées (type_of_meal_plan, room_type_reserved, market_segment_type)
    - Peu de preprocessing nécessaire
- **Lien** : kaggle.com/ahsan81/hotel-reservations-classification-dataset