Analyse et Nettoyage des Données Bottleneck

Huang Nicolas

Développer IA OpenClassRoom

23/10/2025

Contexte

- ***Sources de données : ERP, site web, table de liaison.***

- ***Problèmes :***

  - ***Références produits incohérentes***
  - ***Doublons et valeurs manquantes***
  - ***Formats prix et stock incorrects***
  - ***Impact : indicateurs CA, marge et stock non fiables.***

Objectifs du projet

- ***Nettoyer et agréger les données.***
- ***Produire des indicateurs fiables pour le CODIR : CA, marge, stock.***
- ***Visualiser graphiquement les analyses (Top 50, histogrammes, heatmap)***
- ***Préparer l’industrialisation via pipeline ETL.***

Méthodologie générale

- ***Phase 1 : Agrégation – merge des fichiers ERP, Web et Liaison***
- ***Phase 2 : Nettoyage – suppression des doublons, conversion numérique, flag produits à vérifier***
- ***Phase 3 : Analyse – calcul CA, marge, rotation stock, détection outliers***
- ***Phase 4 : Visualisation – graphiques Matplotlib & Plotly, synthèse tables***

Normalisation des identifiants

```
def norm(x):
    return str(x).strip().upper()

erp['ref_erp_norm'] =
erp['product_id'].astype(str).apply(norm)
web['ref_web_norm'] =
web['sku'].astype(str).apply(norm)
```

1. strip() enlève espaces inutiles
2. upper() uniformise casse
3. Facilite la correspondance entre ERP et Web

Suppression des doublons

```
erp = erp.drop_duplicates
(subset=['ref_erp_norm'])
web = web.drop_duplicates
(subset=['ref_web_norm'])
liaison = liaison.drop_duplicates
(subset=['ref_web_norm', 'ref_erp_norm'])
```

1. Évite que des produits identiques soient comptés plusieurs fois.
2. Garantit un merge sécurisé et des calculs fiables.

Merge sécurisé

```
full = web.merge(liaison, how='outer',
on='ref_web_norm') \
        .merge(erp, how='left',
on='ref_erp_norm')
```

1. outer merge : conserve tous les produits Web même sans ERP
2. left merge : ajoute données ERP quand correspondance existe
3. Résultat : fichier complet pour analyses et flag produits à valider

Conversion numérique et nettoyage

```
full['price_num'] = pd.to_numeric(full['price'],errors='coerce')

full['stock_quantity_num'] = pd.to_numeric(full['stock_quantity'],
errors='coerce').fillna(0).astype(int)

full['purchase_price_num'] =
pd.to_numeric(full['purchase_price'], errors='coerce')

full['a_valider_liaison'] = full['ref_erp_norm'].isna()
```

- pd.to_numeric(..., errors='coerce')

➤ Convertit les valeurs texte en nombre, remplace les erreurs par NaN

- .fillna(0).astype(int)

➤ Remplace les valeurs manquantes par 0 et force un entier pour les stocks.
Permet d'éviter les erreurs dans les calculs de rotation.

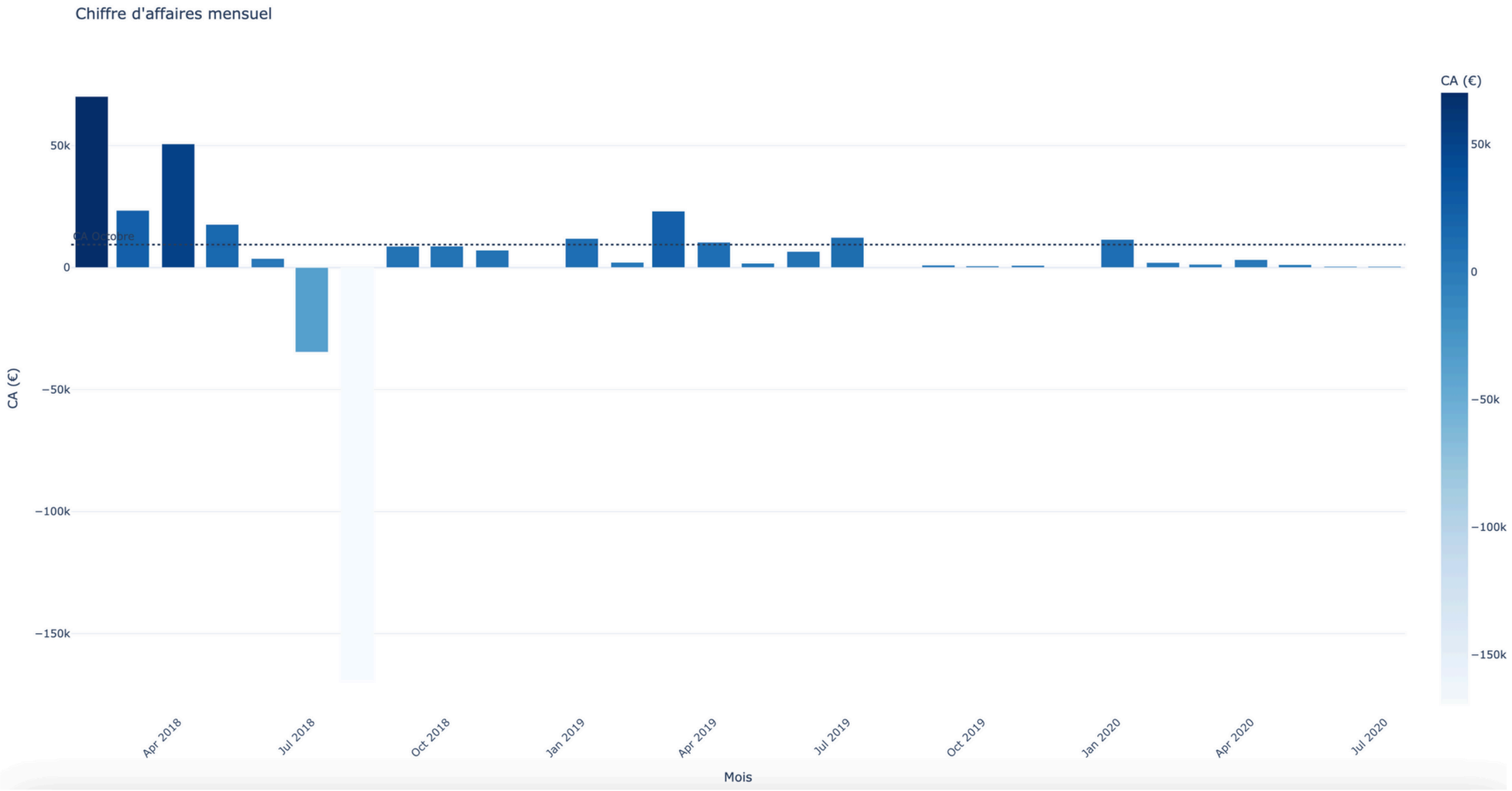
- full['a_valider_liaison'] = full['ref_erp_norm'].isna()

➤ Crée un flag pour les produits présents sur le site Web mais absents de l'ERP.
Sert à isoler les anomalies de correspondance.

Calcul des indicateurs

- ***CA produit : total_sales_value = total_sales * price_num***
- ***Marge : (price_num - purchase_price_num)/price_num***
- ***Rotation stock : total_sales_value / stock_quantity_num***
- ***Mois de stock : stock_quantity_num / total_sales_value***
- ***Argument : indicateurs clés pour prioriser actions business***

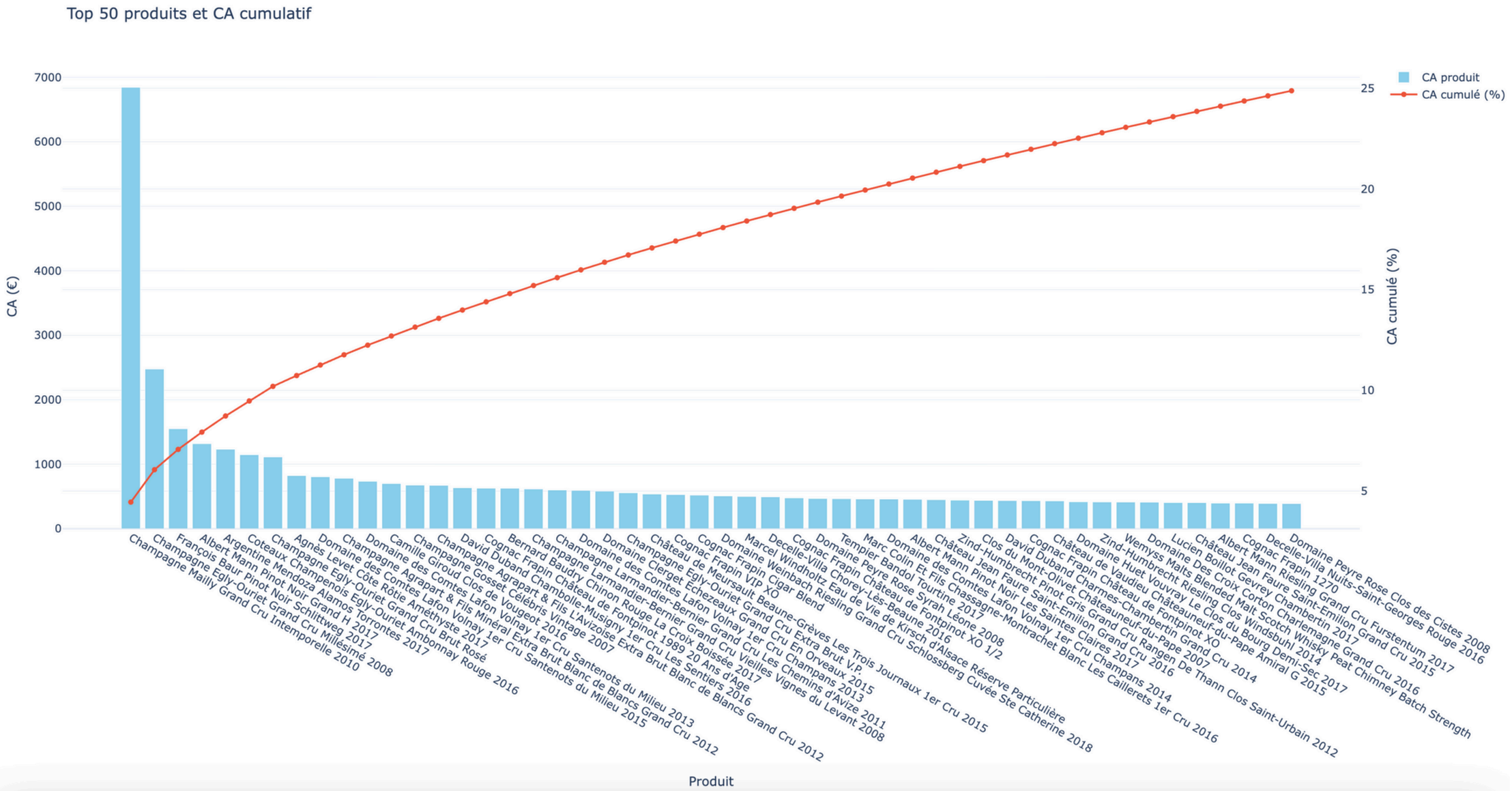
Chiffre d’affaires



CA total : 153 748,10 €

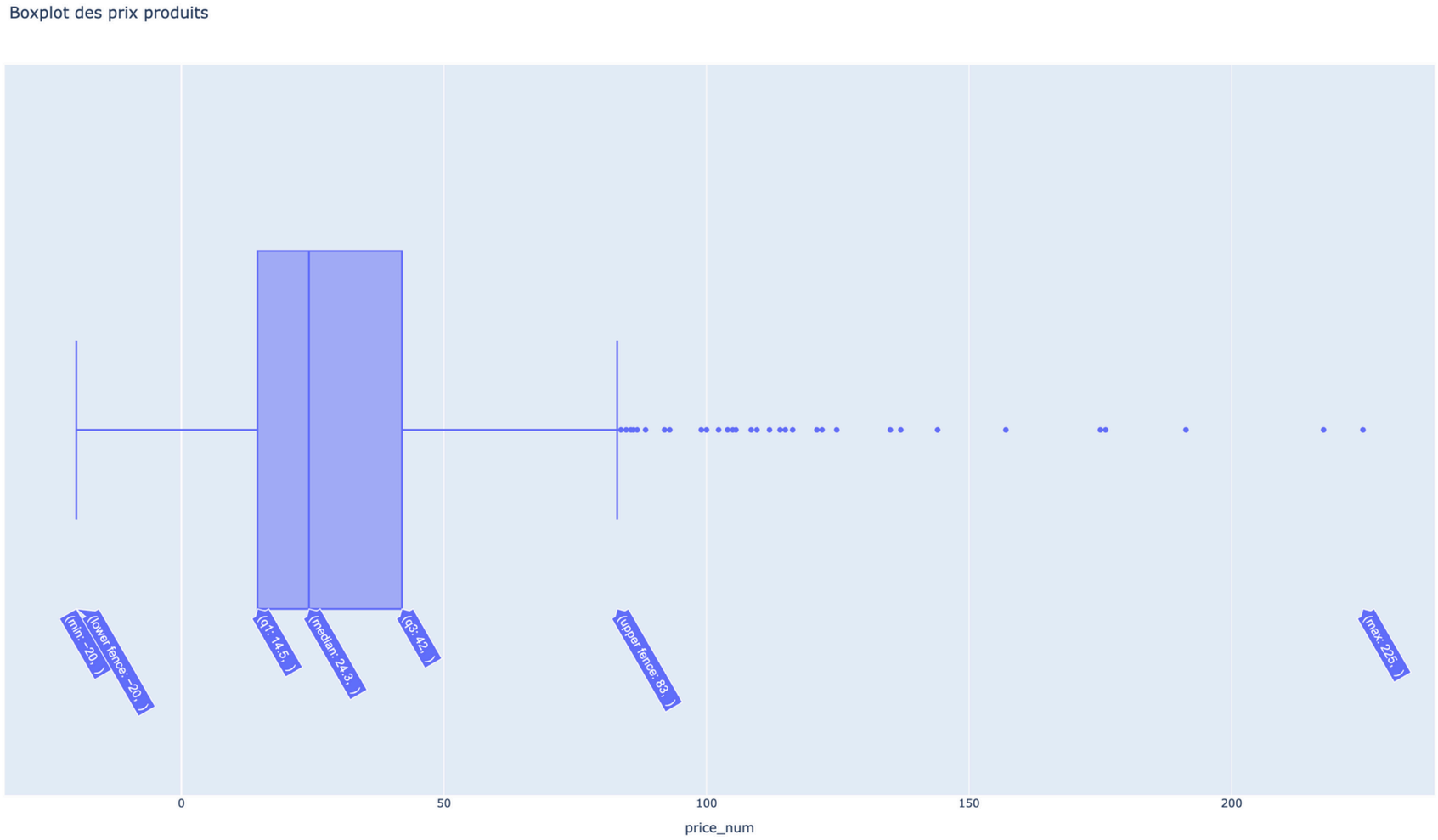
CA octobre : 4 693,50 €

Concentration du CA (Pareto)



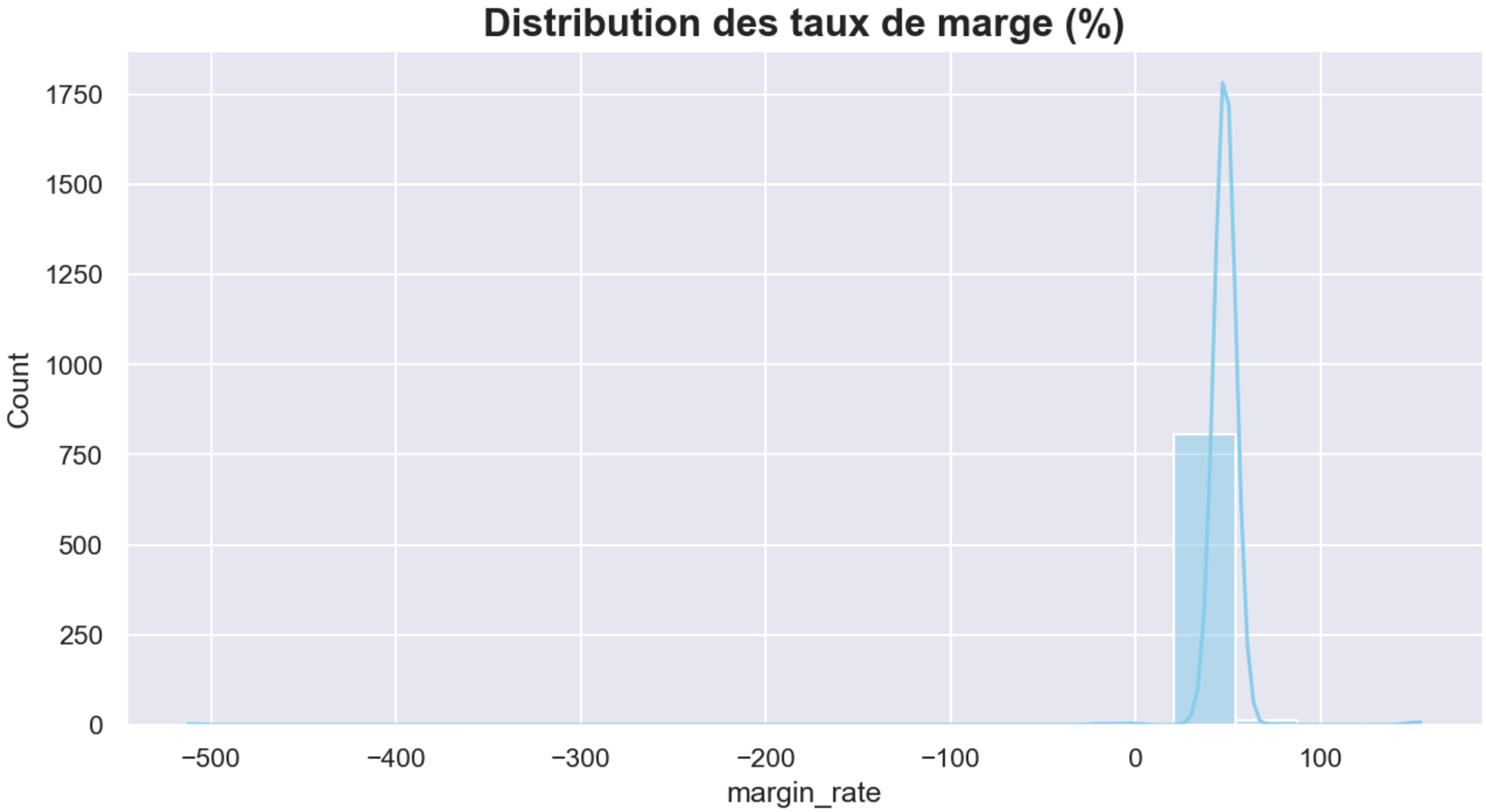
- 420 produits = 80 % du CA
- Graphique Top 50 produits : barre CA + ligne CA cumulé (%)
- Insight : forte concentration → identifier produits stratégiques

Analyses complémentaires
CA, quantités, stocks, taux de marge et correlations



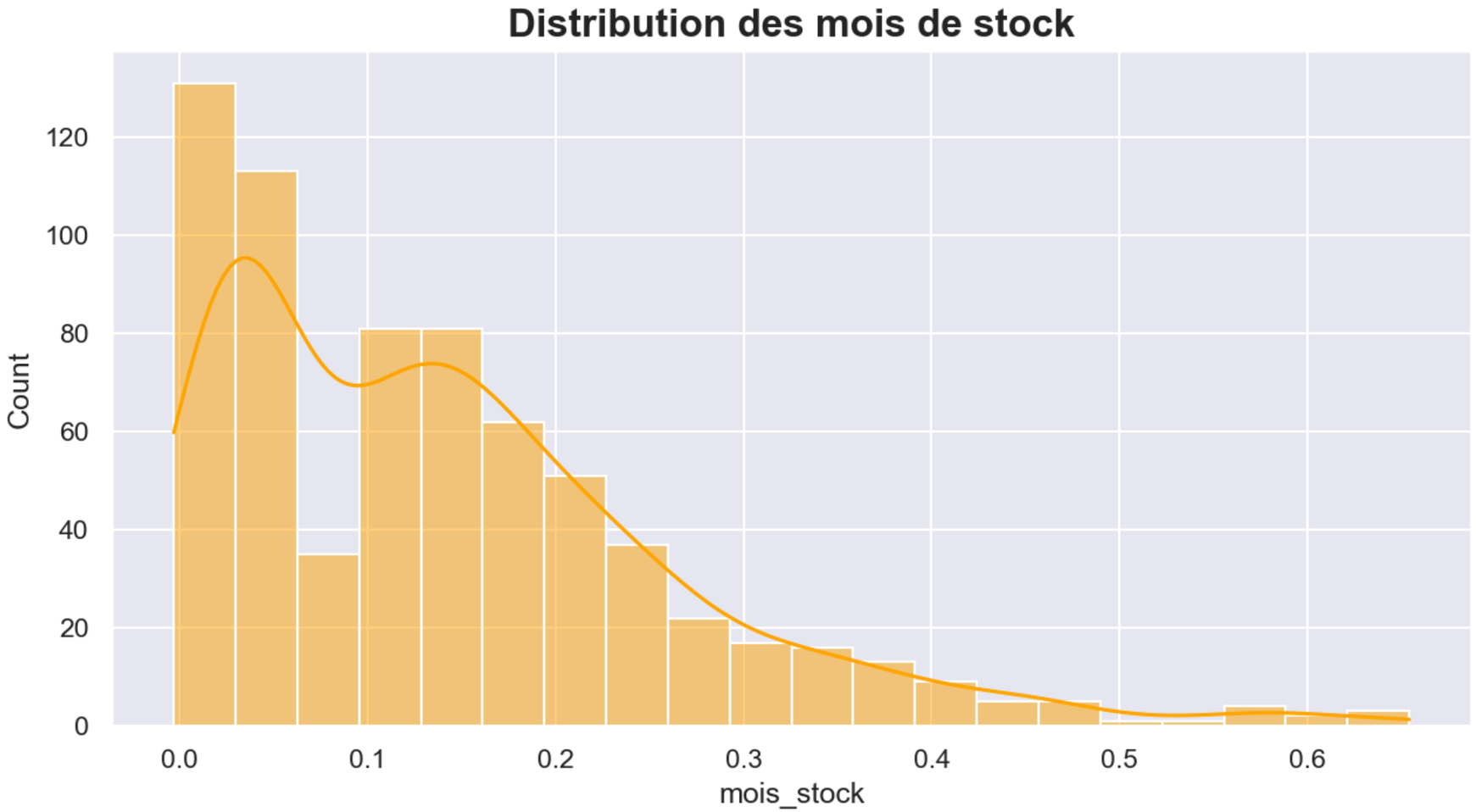
- La boîte montre la répartition normale des prix (Q1–Q3).
- La ligne centrale = prix médian (~25 €).
- Les points à droite représentent les prix aberrants détectés (hors plage normale).
- Environ 36 produits présentent des valeurs incohérentes

Distribution marges



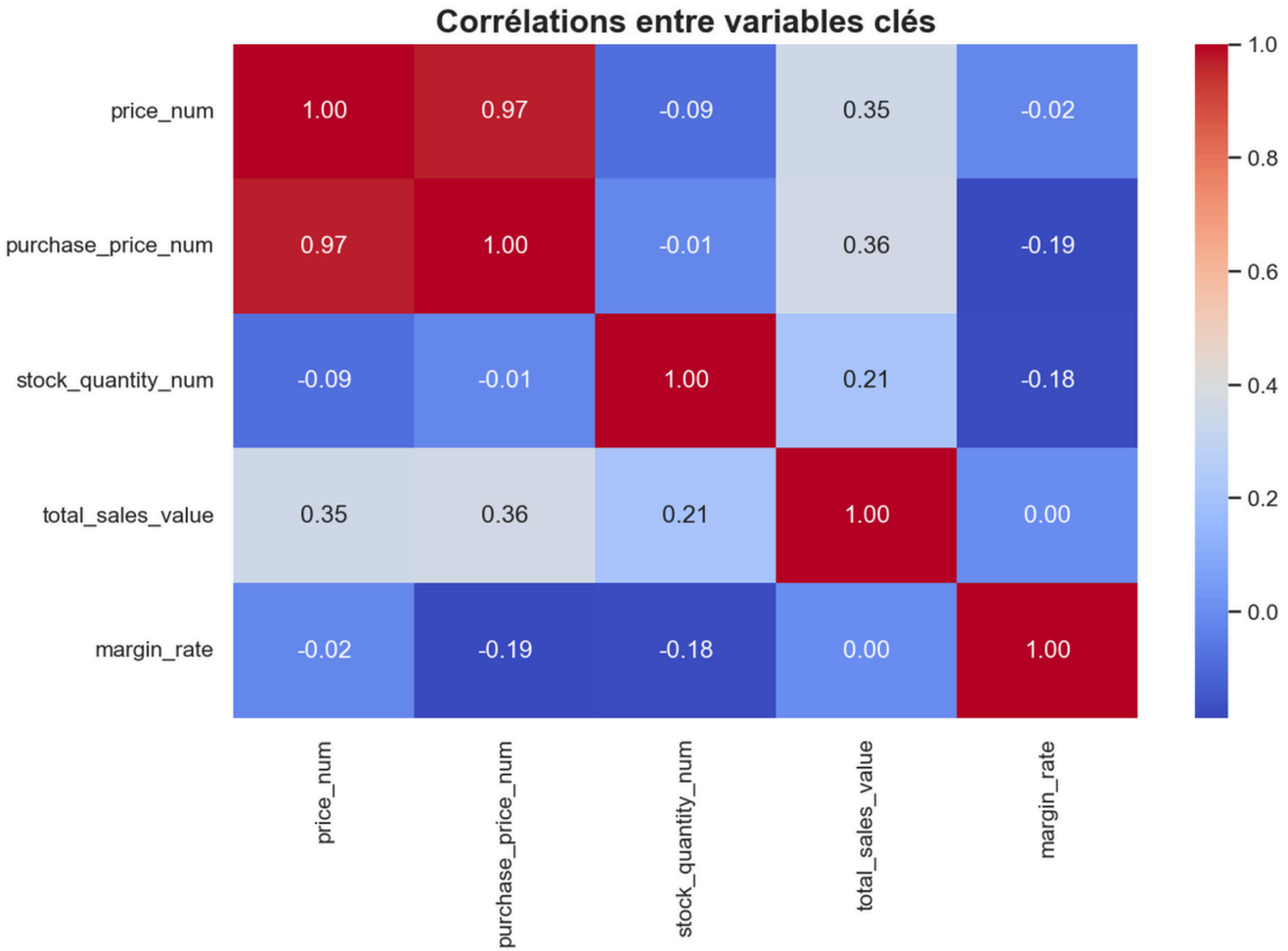
- Histogramme distribution des marges
- Certaines références vendues à perte
- Insight : identifie produits à faible marge pour stratégie prix

Stock critique



- Produits <1 mois : 689
- Tableau exemples stock critique : post_title, stock_quantity, total_sales, mois_stock
- Graphique histogramme mois de stock
- Insight : prioriser réapprovisionnement et gestion des ruptures

Corrélations



- Heatmap corrélations : prix, purchase_price, stock, CA, marge
- Analyse :
- Prix ↔ ventes faible
- Stock ↔ ventes positif mais saturé
- Insight : ajuste prix et stock pour performance commerciale

Synthèse et préconisations

- ***CA total : 153 748 €***
- ***Stock critique : 689***
- ***Prix aberrants : 36***
- ***Produits Web sans ERP : 0 %***

- ***Préconisations :***
- ***Standardiser ERP/Web***
- ***Pipeline ETL et contrôle qualité automatique***
- ***Optimiser marges et rotation stock***

- ***Conclusion : indicateurs fiables pour décisions stratégiques***

Merci de votre écoute

Huang Nicolas

Développer IA OpenClassRoom

23/10/2025