Optimisez la gestion & nettoyez les données du stock d'une boutique BottleNeck

Fine wine spirit

[Felizarda LANDO]

[BI Analyst]

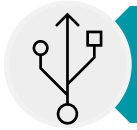
[20/01/2025]

Analyses Exploratoires des Données

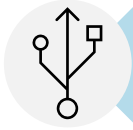
Datasets

Traitement réalisé sur Notebook

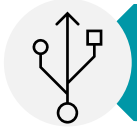
import pandas as pd/ import plotly.express as px



Extraction de l'ERP (référence produit, prix et l'état du stock),



Extraction du site Web (SKU, quantités vendues, description des produits, etc.)



Extraction table de liaison qui permet de lier les références entre la base de données Wordpress et l'extraction de l'ERP de l'entreprise.

Caractéristiques analysées

Caractéristiques	Df_erp	Df_web	Df_liaison
▪ **Structure:** Colonnes (Variables) ,Lignes (Observations). df.shape()	825 observations 6 colonnes	1513 observations 29 colonnes	825 observations 2 colonnes
▪ **Type de données:** df.dtypes()	Int (3), float (2) et objet (1)	float (2), objet (12) et datetime (2)	Objet, Int
▪ **Valeurs non nulles :** df.info()	825 non-null	sku 714 non-null Total_sales 714non-null	Product_id 825 non-null Id_web 734 non-null
▪ **nombre de valeur présente dans chaque colonne:** df.count()	825	- sku 1428 - total_sales 1430	825
▪ **Vérification si doublons:** df[nom colonne].duplicated()	product_id: Absence de doublons	Nombre de doublons : 798	'id_wyeb': Absence de doublons
▪ **Valeurs distinctes:** df_erp['stock_status'].unique()	['instock' 'outofstock']		.
▪ **Moyenne, médiane, minimum, maximum, écart-type:** (df.describe())	(Cf page 5)	(Cf page5)	(Cf page5)

- Traitement realisé df_erp

Traitement realisé

- Identification des erreurs dans les données (erreurs de saisie, etc.).
- Vérification si présence de lignes doublons colonne product_id
- Affichage des valeurs distinctes de la colonne stock_status:
- Creation de la une colonne "stock_status_2:
- Identification des lignes en écart
- Correction des données incohérentes en les supprimant

•Analyse exploratoire de chaque variable du fichier erp.xlsx

df_erp.describe() qui fournit comme information les statistiques sur les colonnes numériques

	product_id	onsale_web	price	stock_quantity	purchase_price
count	823.000000	823.000000	823.000000	823.000000	823.000000
mean	5164.300122	0.867558	32.187303	21.637910	16.941130
std	902.838434	0.339177	26.736076	21.937927	14.575386
min	3847.000000	0.000000	-20.000000	-10.000000	2.740000
25%	4349.000000	1.000000	14.500000	7.000000	7.575000
50%	4908.000000	1.000000	24.300000	18.000000	12.710000
75%	5805.500000	1.000000	41.900000	30.000000	22.015000
max	7338.000000	1.000000	225.000000	145.000000	137.810000

On constate:

- **Prix avec valeur négative →** Il faut les corriger
- **Forte dispersion des prix d'achat →** Vérifier si certains prix anormalement élevés ou faibles.
- **Présence de stocks négatifs →** Problème à corriger

- Analyse de la variable Prix, de la variable STOCK, de la variable ONSALE_WEB, de la variable prix d'achat

verification	prix	stock	Onsale_web	Prix_d'achat
Colonne	Name: price, dtype: float64	Name: stock_quantity, dtype: int64	Name: onsale_web, Length: 823, dtype: int64	Name: purchase_price, Length: 823, dtype: float64
prix négatifs ou nuls	Trois prix négatif détecté: -20,-8,-9,1 0 articles avec un prix non renseigné	/	/	/
Affiche age du max et min:	Le prix minimum est : -20.0 Le prix maximum est : 225.0	La quantité minimum est : -10 La quantité maximum est : 145		Le prix minimum est : 2.74 Le prix maximum est : 137.81
stocks inférieurs à 0	/	Stock_quantity : -10;-1	/	/
Action pour corriger la ou les données incohérentes	Ici il a été décidé de modifier les valeurs négatives en valeurs positivesdf_erp['price'] = df_erp['price'].abs() # La méthode abs() rend les valeurs positives	les supprimer	Supprimer la colonne "stock_status_2" car e redondante avec la colonne "stock_status	/

- Analyse exploratoire du fichier web.xlsx et du fichier liaison.xlsx

Identification des colonnes à conserver **df_web**(cf note book)

Visualisation des valeurs de la colonne **sku** qui ne respecte pas la règle de codification(NaN)

Identification des anomalies

Anomalie	Causes possibles
Présence de NaN	- Données manquantes ou non renseignées - Problème d'importation ou de fusion de données
Format incorrect (caractères spéciaux, lettres, tirets, espaces, etc.): 13127-1, bon-cadeau-25-euros	- Mauvaise saisie manuelle - Règle de nommage spécifique pour certains produits
Numéro trop court ou trop long: 38, 1982145623	- Problème lors de la génération automatique des codes - Erreur humaine lors de la saisie
Présence de doublons avec variations: 13127, 13127-1	- Code de base correct mais ajout de suffixes pour différencier des variantes - Mauvaise gestion des références produits

- Solutions pour améliorer les données dans nos systèmes.

Type d'anomalie	Action corrective
(valeurs manquantes) NaN	- Identifier les produits concernés et récupérer l'information - Remplir les valeurs manquantes en se basant sur les catalogues existants ou les bases de données métier
Format incorrect (sku avec caractères spéciaux, lettres, tirets, etc.)	- Définir une règle stricte de nommage et la communiquer aux équipes - Mettre en place une validation automatique lors de la saisie
Numéro trop court ou trop long	- Vérifier si l'erreur provient d'une mauvaise saisie ou d'un problème d'importation - Uniformiser la longueur des sku (ex. : tous à 5 ou 6 chiffres)
Doublons avec variations (13127, 13127-1)	- Vérifier s'il s'agit du même produit ou d'une variation - Mettre en place un système de gestion des variantes (ex. : attributs séparés au lieu d'un sku modifié)

- Nettoyages des données et Features engineering

- Remarques éventuelles, pièges ou difficultés rencontrées

- Difficultés sur certains choix décisionnelle pour le nettoyage des données
- difficulté avec un code créer qui ne supprimait pas les colonnes voulu(les colonnes contenant uniquement des zéros)
- Erreur de syntaxe (crochet, une virgule…)

Fusion ou consolidations des données

- *Choix des attributs*

Jointure entre df_erp et df_liaison

sur la colonne 'product_id' .

- *Clés utilisées*
- product_id' → elle présente dans les deux datasets et contient des valeurs uniques
- *df_merge = pd.merge(df_erp, df_liaison, how="outer", on="product_id", indicator=True)df_merge[df_merge["_merge"] != "both"]df_merge.drop("_merge", axis=1, inplace=True)display(df_merge)*

Jointure df_merge et df_web

sur la colonne 'id_web' et 'SKU'

Clés utilisés

- *id_web et SKU*
- *df_global = pd.merge(df_merge, df_web, how="left", left_on="id_web", right_on="sku")*

- *Vigilances particulières au cours du traitements:*

-Que l'attribut existe dans les deux datasets et contienne des valeurs qui permettent d'établir une correspondance claire.

-Bien choisir le type de jointure en fonction du besoin : ex:Inner Join , outerjoin

- *Difficultés ou pièges rencontrés*

-Pour effectuer les jointures. La première jointure assez compliqué a réaliser pour ma part

-Erreurs de syntaxe

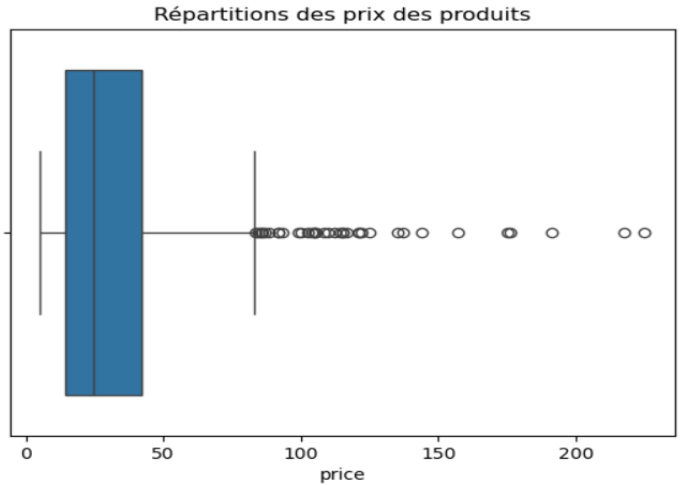
-Colonnes dupliquées

Analyses univariées du prix

- *Méthodes statistiques employés*

Création d'une Boite à moustache de la répartition des prix grâce à Pandas

- *Graphique avec commentaire des résultats*



Les prix sont concentrés dans une plage étroite

la médiane(prix median 24,3€) est plus proche du bas de la boîte → Il y a plus de produits à bas prix qu'à prix élevé.

tout les outliers (prix entre 83 et 225€) sont au-dessus

borne inférieure : Premier quartile (Q1,prix 14,50), soit les 25% des prix les plus bas.
Borne supérieure : Troisième quartile (Q3, prix 41,93), soit les 75% des prix les plus bas.
Ligne à l'intérieur :50% des produits ont un prix en dessous et 50% au-dessus). Les moustaches (whiskers)

Limites éventuelles de l'analyse

→ Un boxplot ne permet pas de voir **s'il existe plusieurs modes** (pics de concentration dans la distribution des prix).

Cas ci-dessous on vend des vins d'entrée de gamme à 5 € et des vins premium à plus de 100€, un boxplot ne permet pas d'identifier ces **deux catégories distinctes**.

→ Le boxplot identifie des valeurs comme **outliers** uniquement sur la base de la règle statistique des 1.5×IQR (Intervalle Interquartile).

Mais certains outliers peuvent être normaux ! Si une bouteille de vins grands cru est plus cher que les autres, ce n'est pas une forcement une erreur mais une caractéristique du marché.

→ Perte d'information car **boxplot ne montre pas la fréquence des valeurs**.
On sait où sont les quartiles, mais on ne voit pas **combien** de produits ont un prix proche de la médiane ou des extrêmes.

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

Méthodes statistiques employés

- *Calcule du CA global → 153423,10€*
- *Calcul du total de produit vendus → 6068 bouteilles de vins et spiritueux vendus*
- *Calcul somme des stocks totaux* → 16481 bouteilles en stocks
- *Calcul du taux de marge et corrélation* → taux de marge min:22,8% / Taux de marge max:47,75%
- *Valorisation des stocks → 480 414 € grande quantité de capital immobilisé en stock.*

Limites éventuelles de l'analyse:

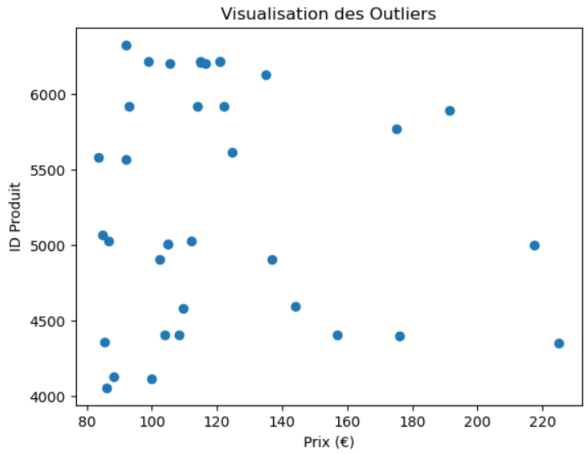
- Certaines bouteilles peuvent avoir une saisonnalité qui n'est pas reflétée dans les chiffres globaux.

- Un CA élevé ne signifie pas nécessairement une forte rentabilité. Il faudrait aussi analyser les coûts d'achat et les frais liés au stockage.

- L'analyse du taux de marge ne donne pas d'indication sur la répartition par gamme de produits.

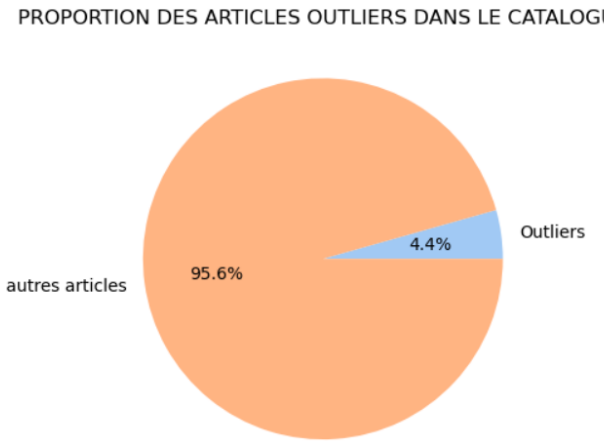
- Manque d'analyse des comportements clients. Une analyse par catégorie de clients, canal de vente ou tendance du marché pourrait apporter plus de précisions.

- *Graphique avec commentaire des résultats*



LES OUTLIERS

Sur 36 outliers La grande majorité des prix se situe entre 83 et 120€ seul une dizaine de bouteille ont un prix au dessus de 120€ et seulement 5 bouteilles au dessus de 150€ et 2 coute plus de 200€



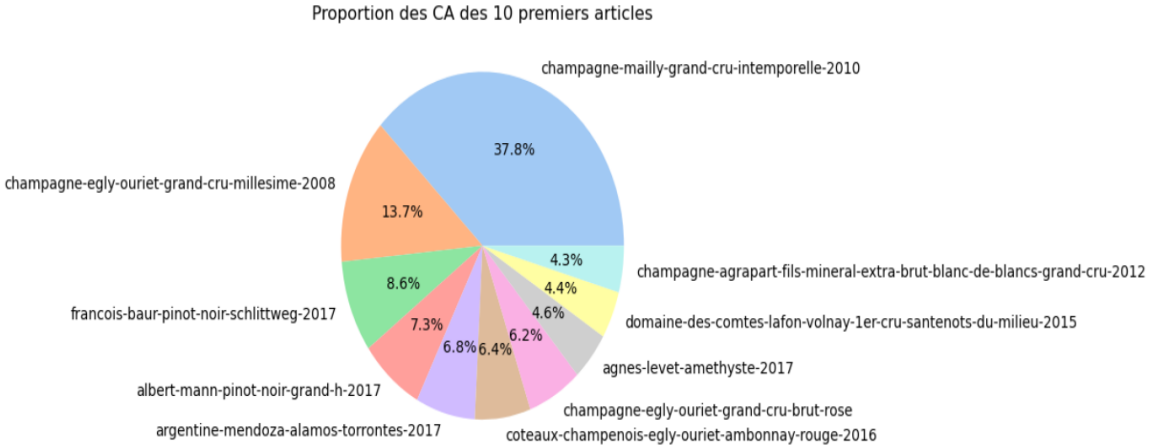
% DANS LE CATALOGUE

Les outliers represente une toute petite partie des bouteilles de vins du catalogue 4.4%

TOP 20 CA

le Champagne mailly grand cru intemporelle 2010 réalise de loin le meilleurs CA 6844€

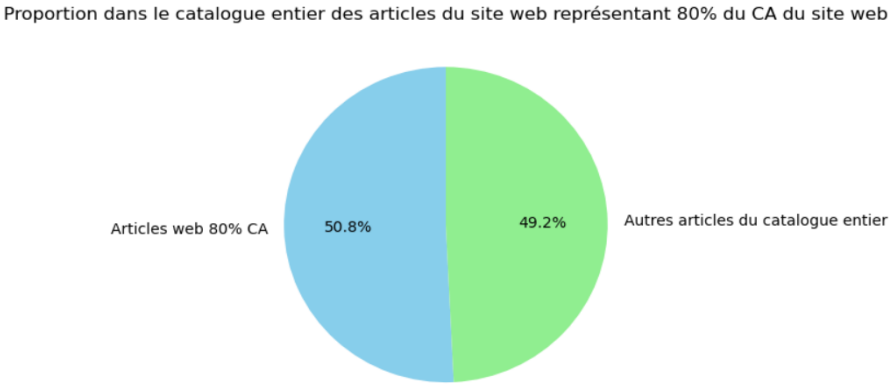
Suivi du champagne egly grand cru millesime 2008 2475€ de CA
Puis les CA realisés pour 5 autres vins se situe entre 15059 et 1113€



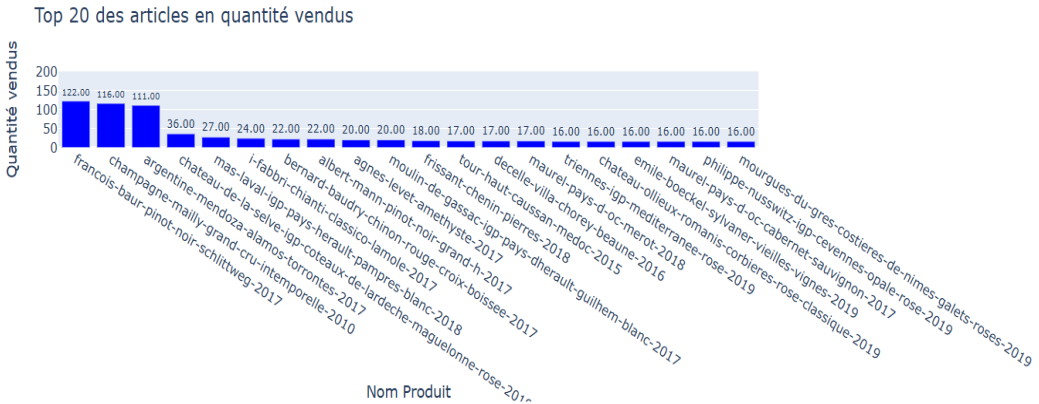
% 10 PREMIERS CA

le Champagne mailly grand cru intemporelle represente 37,8%, plus 1/3 du CA globale

%ARTICLES REPRESENTANT 80% DANS LE CATALOGUE ENTIER



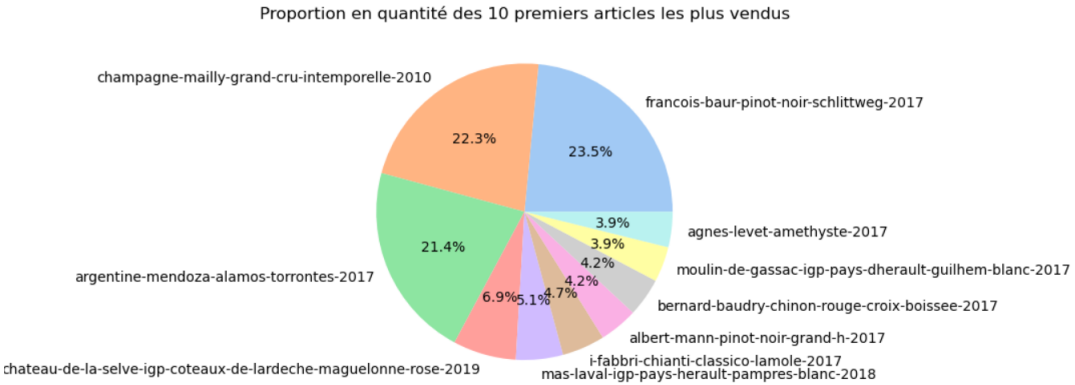
Les articles web représentent 80% du CA représente plus de la moitie du CA catalogue entier(50,8%)



TOP 20 QUANTITÉS VENDUS

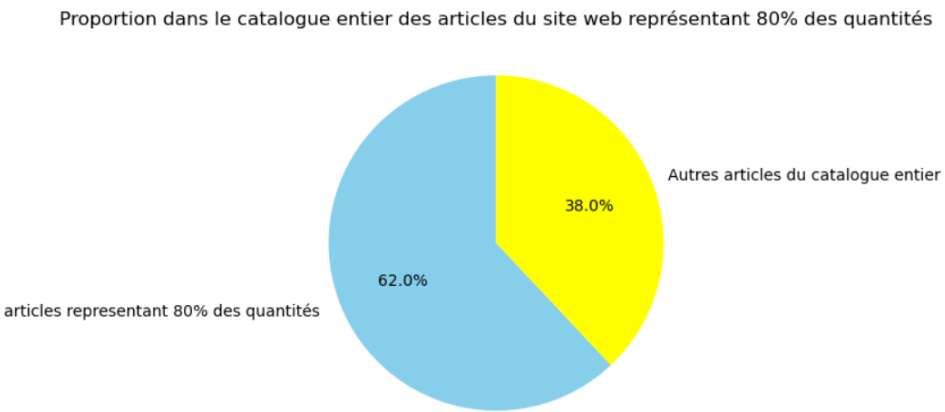
En quantitité vendu trois vins se partage le poduim francois baur pinot avec 122 unités

Suivis par **champagne mailly grand cru 116 unités** vendus et en troisième position argentine mendoza 2017, 111 unités vendus



TOP10 % EN QUANTITE VENDUS

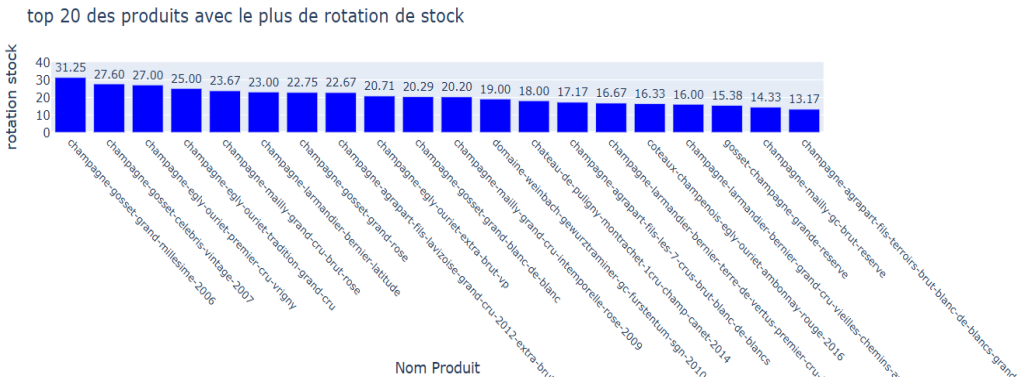
Proportionnellement le champage mailly 22,3%, le françois baur 23,5% et le argentine mendoza 21,4% constituent a eux trois les ¾ des produits les plus vendus



% DANS CATALOGUE DES ARTICLES WEB REPRESENTANT 80% DES QUANTITES

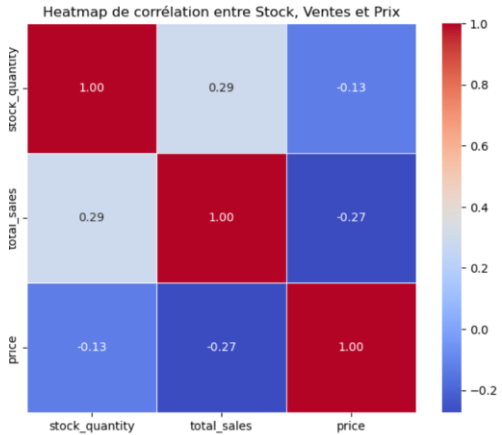
Les articles du site web, représentant 80% des quantités vendus représentent 62% des quantités vendus du catalogue entier

TOP 20 ROTATION DE STOCK



Les champagnes sont la catégorie d'articles qui on a les meilleures rotations de stock. Les stocks sont généralement très bien gérés il dure entre 1 et 2 semaines.

CORRELATION STOCK, VENTE ET PRIX



-0,13 il n'y a pratiquement pas de relation entre la quantité de stock et le prix des produits. Ii se peut que Les produits chers solt stockés en plus petite quantité et inversement.

Une corrélation de 0.29 montre ici qu'il existe une légère tendance à ce que les articles avec une plus grande quantité de stock aient des ventes plus élevées, mais cette relation reste faible.

-0.27 suggère que, en général, lorsque le prix des produits augmente, les ventes totales ont tendance à diminuer légèrement, et vice versa.

Actions pour la suite

Gestion des données

Nettoyer et Préparer les Données Régulièrement.
Catégoriser les produits, ajouter une colonne classe.
Regrouper les ventes par mois trimestre et année.
Former les équipes à l'importance de la qualité des données.
Surveiller en continu la qualité des données via des audits et des outils adaptés.

Travailler sur la sensibilité prix-vente:

Tester des **ajustements de prix dynamiques** sur les bouteilles les moins vendus.
Proposer **des packs ou offres groupées** pour écouler les stocks tout en préservant la marge.

Maximiser les ventes des produits à fort CA:

Le *Champagne Mailly Grand Cru Intemporelle 2010* réalise **37,8% du CA globale**, et d'autres vins premium génèrent un bon CA
Renforcer la mise en avant de ces produits sur le site et en magasin (ex: promotions ciblées, recommandations personnalisées).
Analyser les comportements d'achat : qui achète ces produits et pourquoi ?

Renforcer la stratégie web

Investir davantage dans le e-commerce, publicité ciblée et partenariats avec des influenceurs.

Optimiser la gestion des stocks & rotation:

Ajuster les commandes pour éviter le surstockage des grands crus qui restent longtemps en stock.
Analyser pourquoi certaines bouteilles restent bloqués en stock et envisager des réductions sur ces références pour accélérer leur vente.

Avec ces actions, BottleNeck pourras **augmenter la rentabilité**, **réduire les coûts liés au stock**, et **maximiser les ventes via une meilleure stratégie digitale**.

Point sur les compétences apprises

Qu'est-ce qui s'est bien passé pour vous dans ce travail de nettoyage ?

- Travailler avec un notebook
- Effectuer des analyses exploratoires de fichiers.
- Effectuer des analyses univarié.
- Effectuer les visualisations , les personnaliser.

Qu'est-ce que vous avez trouvé le plus difficile ?

-Coder en python car c'est très différent de SQL,

-Effectuer la jointure entre df_erp et df_liaison , (code partiellement incorrect et pb de duplication des cellules)

-Effectuer les visualisations , les personnaliser m'as pris beaucoup de temps

Sur quelles tâches est-ce que vous pensez avoir besoin de plus d'entraînement ?

-Gagner en aisance pour effectuer des jointures complexe

-Effectuer les visualisations , les personnaliser. Je travail en ce sens .