Optimisez la gestion des données d'une boutique avec Python

Arnaud Golliot
Data Analyst
8 novembre 2024

Contexte de l'étude

Nicolas, mon manager en charge de l'analyse des données d'activité de Bottleneck, souligne les difficultés opérationnelles suivantes :

- ✓ La gestion des stocks est complexe avec des outils artisanaux
- ✓ Les compétence et outils en analyse de données sont déficients

L'objectif est d'établir une première analyse des données relative à l'activité globale de Bottleneck autour des indicateurs suivants : prix de vente, CA, quantités vendues, taux de marge, taux de rotation des stocks. A partir de ces indicateurs, proposer des orientations stratégiques.

Une fois finalisée, cette étude sera présentée en CODIR afin de les éclairer sur les points d'amélioration de l'activité, et valider avec eux la pertinence des orientations stratégiques proposées avant de décider (ou non) de les mettre en œuvre.

Pour effectuer cette analyse, Bottleneck dispose des données brutes suivantes :

- ✓ L'extraction de l'**ERP** (référence produit, prix et l'état du stock)
- ✓ L'extraction de notre **site Web** (SKU, quantités vendues, description des produits, etc.)
- ✓ Une **table de liaison** qui permet de lier les références entre la base de données Wordpress et l'extraction de l'ERP de l'entreprise
  - Les numéros des références ne correspondent pas entre les deux outils.
  - Notre stagiaire a mis à jour cette liste pour toi avec les nouveaux produits.

Description des données

Trois fichiers Excel à analyser : erp.xlsx, web.xlsx, liaison.xls

erp.xlsx	contient le catalogue des produits avec les quantités en stock					
	Données exploitées	**Nom de la variable**	**Description**	**Type**	**Identifiant**	**Observations**
		product_id	Idnetifiant du produit dans l'ERP	Numérique	X	Exemple : 4150
		price	prix de vente TTC du produit (supposé en €)	Numérique		Exemple : 59
		purchase_price	prix d'achat HT du produit (supposé en €)	Numérique		Exemple : 35.45
		stock_quantity	Quantité en stock (en nombre d'unités)	Numérique		Exemple : 123
		stock_status	Statut du stock	Discrète		Deux valeurs : "instock" ou "outofstock"
	Identifiant	product_id				
	Etendue temporelle	Sans objet, référentiel non historisé				
	Commentaire	Si le stock est à zéro, alors la variable stock_status est supposée contenir la valeur "outofstock". Si le stock > 0, alors la variable stock_status est supposée contenir la valeur "outofstock"				

web.xlsx	Contient les achats effectués sur le site de vente en ligne					
	Données exploitées	**Nom de la variable**	**Description**	**Type**	**Identifiant**	**Observations**
		sku	Code article sur le web	Numérique	X	Exemple : 1366
		tax_status	Type de page consultée	Discrète	X	Deux valeurs possibles : "taxable" ou vide
		total_sales	Quantité vendue	Numérique		Exemple : 10 (unités)
		product_type	Catégorie de produit	Discrète		6 valeurs possibles : "Vin", "Champagne", "Cognac", "Whisky", "Gin", "Huile d'olive"
		post_title	Dénomination de l'article	Discrète		Exemple : "Champagne Mailly Grand Cru Intemporelle 2010"
	Identifiant	sku, tax_status				
	Etendue temporelle	du 08/02/24 au 20/07/24, soit 3 ans et 5 mois				
	Commentaire	Le fichier contient deux lignes par article : la page descriptive de l'article, et la photo de l'article				

Liaison.xlsx	Contient les liens entre code article dans la catalogue des produits et le code article enregistré sur le site de vente en lignes					
	Données exploitées	**Nom de la variable**	**Description**	**Type**	**Identifiant**	**Observations**
		id_web	Code article sur le web	Numérique	X	Exemple : 1366 Identique au sku enregistré sur le site de vente
		product_id	Type de page consultée	Numérique	X	Exemple : 4150 Identique au product_id enregistré dans l'ERP
	Identifiant	id_web, product_id				
	Etendue temporelle	Sans objet, référentiel non historisé				
	Commentaire	Ce fichier permet d'établir la liaison entre les fichiers erp.xlsx et web.xlsx				

Analyses Exploratoires et nettoyage des données

Trois datasets issus du chargement des trois fichiers Excel : ERP, WEB et LIAISONS

Dataset : ERP		Nombre de lignes avant nettoyage	825
Test	**Nb lignes concernées**	**Action corrective**	**Nb lignes conservées**
product_id non renseigné	0	Aucune	825
recherche de doublons sur product_id	0	Aucune	825
prix de vente < 0	3	Suppression des produits concernés	822
prix d'achat < 0	0	Aucune	822
prix de vente < prix d'achat	4	Aucune On **conserve** ces lignes (marge négative)	822
stock < 0	2	Suppression des produits concernés	820
stock différent de zéro et statut de rupture de stock positionné à "outofstock"	1	Aucune	819
		Nombre de lignes après nettoyage	**819**

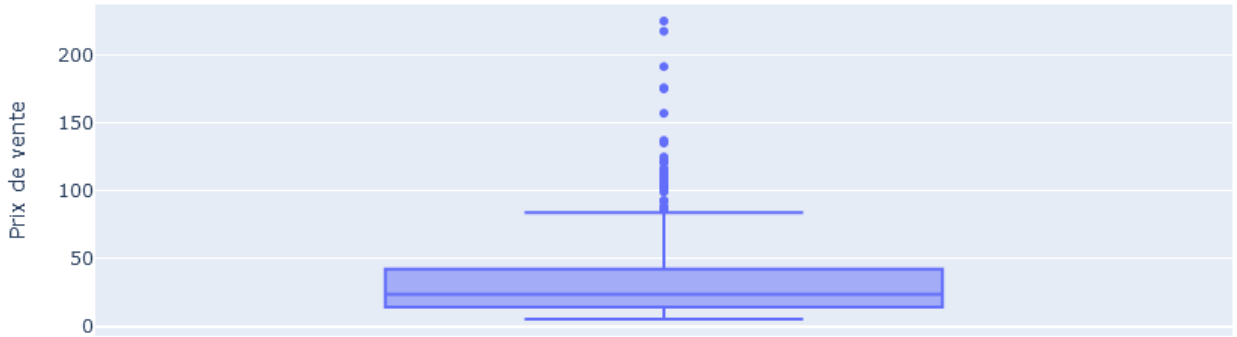
Dataset : WEB		Nombre de lignes avant nettoyage	1430
Test	**Nb lignes concernées**	**Action corrective**	**Nb lignes conservées**
sku non renseigné	2	Aucune	1428
sku non numérique	4	Aucune	1424
doublons sur sku	712	Suppression des lignes dont tax_status est vide	712
		Nombre de lignes après nettoyage	**712**

Dataset : LIAISON		Nombre de lignes avant nettoyage	825
Test	**Nb lignes concernées**	**Action corrective**	**Nb lignes conservées**
id_web (sku) non renseigné	91	Suppression des liaisons concernées	734
id_web non numérique	3	Suppression des liaisons concernées	731
doublons sur id_web	0	Aucune	731
		Nombre de lignes après nettoyage	**731**

Fusion des données

Fusion des 3 Datasets ERP, LIAISON, et WEB		
Nombre de lignes des 3 datasets	ERP	819
	LIAISON	731
	WEB	712
Critères de fusion	ERP et LIAISONS	erp.product_id liaison.product_id
	LIAISON et WEB	liaison.product_id web.sku
Nombre de lignes communes		710
Observations	Les lignes en écart correspondent : >> soit à des produits dans ERP non commandés dans WEB >> soit à produits commandés dans WEB sans liaison dans ERP suite à suppression dans ERP de références dans la phase de nettoyage	
Action	Seules les lignes communes sont conservées, soit 710	

Analyses univariées du prix (répartition)

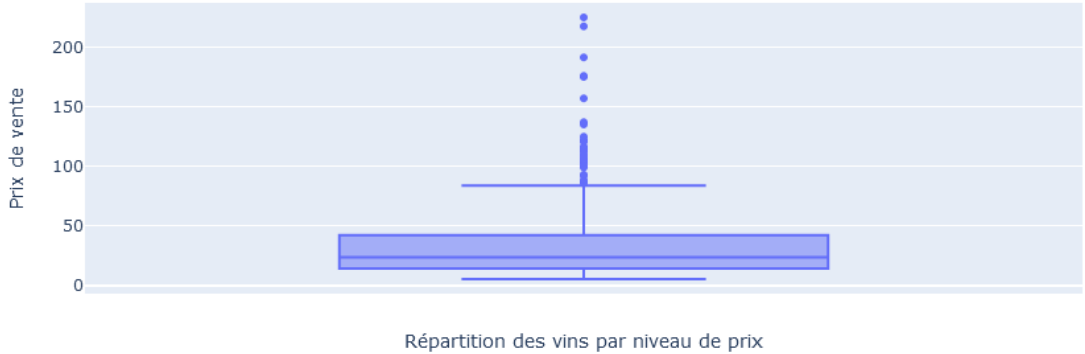


Répartition des vins par niveau de prix

Analyses univariées du prix (moyenne, z-score)

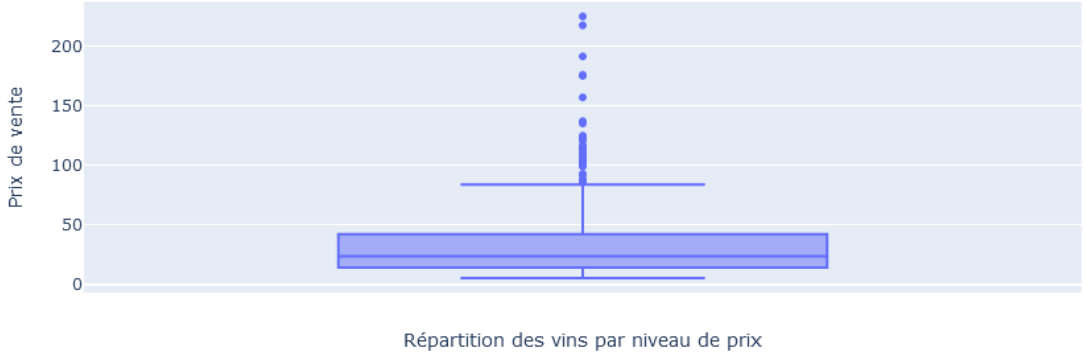
	price
count	710.000000
mean	32.275986
std	27.651235
min	5.200000
25%	14.012500
50%	23.400000
75%	41.950000
max	225.000000

	product_id	price	z_score
24	4402	176.0	5.197743
33	4406	157.0	4.510613
168	4904	137.0	3.787318
198	5001	217.5	6.698580
224	5917	122.0	3.244847
244	5612	124.8	3.346108
245	6126	135.0	3.714988
257	5892	191.3	5.751064
268	6216	121.0	3.208682
269	6213	121.0	3.208682
278	6202	116.4	3.042324
299	5767	175.0	5.161578
547	4352	225.0	6.969816




```
Nombre d'articles pour lesquels le z_score est supérieur à 3 = 13
```

Analyses univariées du prix (médiane, IQR)



	price
count	710.000000
mean	32.275986
std	27.651235
min	5.200000
25%	14.012500
50%	23.400000
75%	41.950000
max	225.000000

4.2.2.3 - Calcul de l'intervalle interquartile IQ = Q3-Q1

```
# Calculer IQ
Price_IQ = Price_Q3 - Price_Q1

# Afficher le résultat
print("Ecart interquatile = {}€".format(round(Price_IQ,2)))
```

```
Ecart interquatile = 27.94€
```

```
Nombre de valeurs aberrantes en ddessus du seuil supérieur : 31 observation(s)
```

```
Nombre de valeurs aberrantes en dessous du seuil inférieur : 0 observation(s)
```

```
% de ces outliers par rapport à la totalité du catalogue : 4.37%
```

4.2.2.3 - Calcul des seuils inférieurs à Q1 et supérieurs à Q2 ¶

```
# Calcul des seuils inférieurs et supérieurs de prix
# Dans l'analyse interquartile standard :
# - le seuil inférieur est le prix minimal à condition que ce minimum ne soit pas inférieur à Q3-1.5IQ, à défaut ce sera Q1-1.5xIQ
# - le seuil supérieur est le prix maximal des prix à condition que ce maximum ne soit pas supérieur à Q3+1.5xIQ

# Autrement dit, le seuil standard dans l'analyse univariée, c'est une valeur qui est
#- soit supérieure à 1.5 fois l'écart intercatile IQ = Q3-Q1
#- soit inférieure à 1.5 fois l'écart intercatile IQ = Q3-Q1
# le 1,5 est une convention empirique

# prix seuil inférieur = Max(Min(prix); Q1-IQ*1.5)
prix_seuil_inferieur = max(df_web_erp['price'].min(), (Price_Q1 - (1.5*Price_IQ)))
print("seuil inférieur de prix = {}€".format(round(prix_seuil_inferieur,2)))

# prix seuil supérieur = Min(Max(prix); Q3+IQ*1.5)
prix_seuil_superieur = min(df_web_erp['price'].max(), (Price_Q3 + (1.5*Price_IQ)))
print("seuil supérieur de prix = {}€".format(round(prix_seuil_superieur,2)))
```

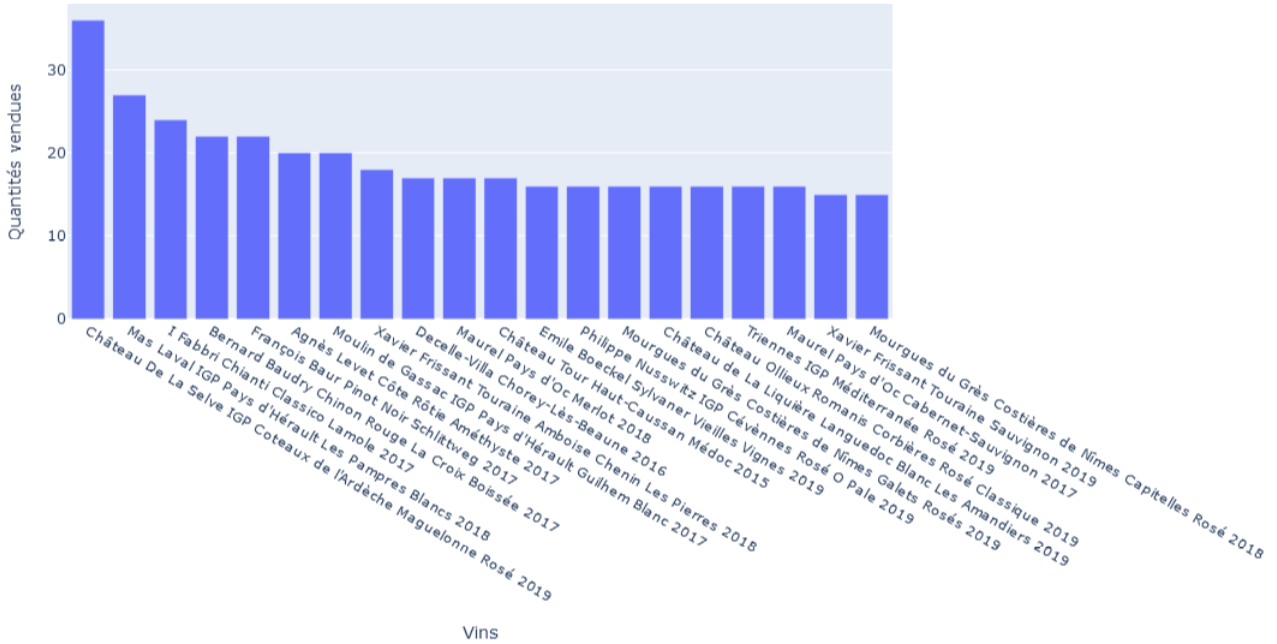
```
seuil inférieur de prix = 5.2€
seuil supérieur de prix = 83.86€
```

Calcul et analyse univariée du CA (20/80)

Nombre d'articles qui représentent 80% du CA : 660 sur 710 soit 92.96%

CA des 80% d'articles vendus sur le CA total : 113 995.0€ sur 142 791.9€ soit 79.83%

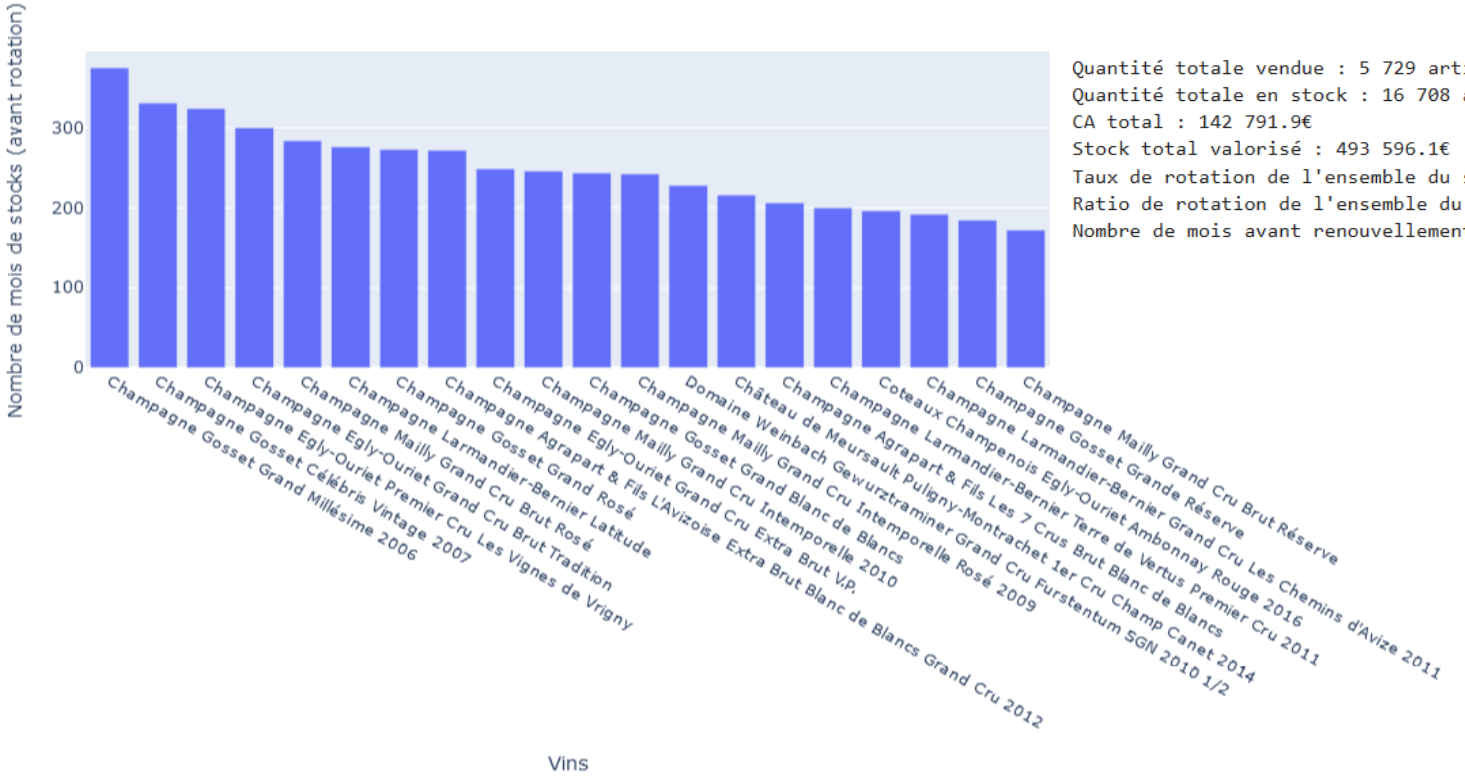
Analyse univariée des quantités vendues



Nombre d'articles qui représentent 80% des quantités vendues : 570 sur 710 soit 80.28%

Quantité des 80% d'articles vendus sur la quantité totale vendue : 3 789 sur 5 729 soit 66.14%

Analyse univariée des stocks incluant le calcul du taux de rotation (flop 20, rotation globale)

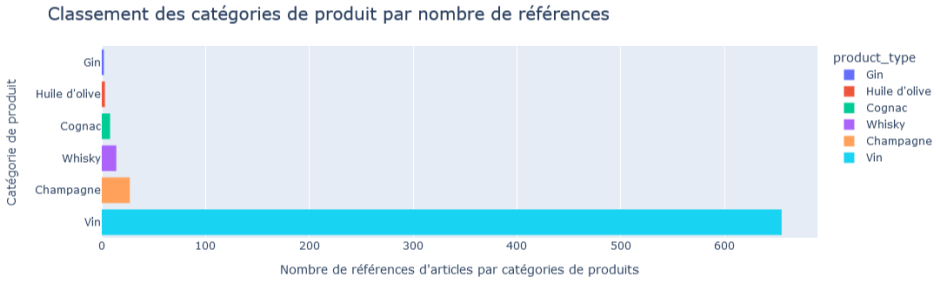


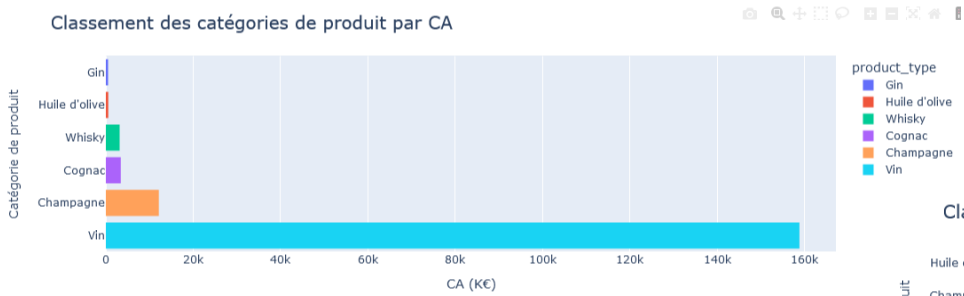
Quantité totale vendue : 5 729 articles
Quantité totale en stock : 16 708 articles
CA total : 142 791.9€
Stock total valorisé : 493 596.1€
Taux de rotation de l'ensemble du stock : 0.29
Ratio de rotation de l'ensemble du stock (en nombre d'années) : 3.46
Nombre de mois avant renouvellement total du stock : 41.52

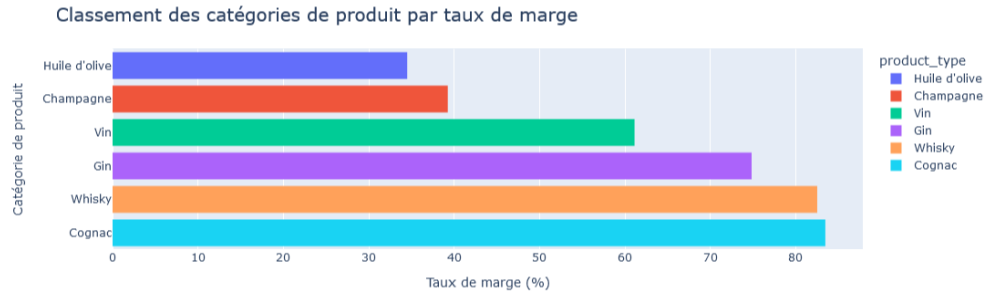
Calcul et analyse univariée du taux de marge

	Article le moins cher	Article le plus cher	Article le moins rentable	Article le plus rentable
Code article	14570	15940	11258	14774
Libellé article	Moulin de Gassac IGP Pays d'Hérault Guilhem Bl...	Champagne Egly-Ouriet Grand Cru Millésimé 2008	Huile d'Olive Extra Vierge Planeta 50cl	Wemyss Malts Single Cask Scotch Whisky Chocola...
Quantité vdue	20	11	7	1
Prix de vente TTC	5.2€	225.0€	13.1€	93.0€
CA	104.0€	2475.0€	91.7€	93.0€
Prix de vente HT	4.33€	187.5€	10.92€	77.5€
Prix d'achat HT	2.74€	137.81€	8.43€	40.49€
Marge	1.59€	49.69€	2.49€	37.01€
Taux de marge	58.03%	36.06%	29.54%	91.41%

Calcul et analyse univariée du taux de marge

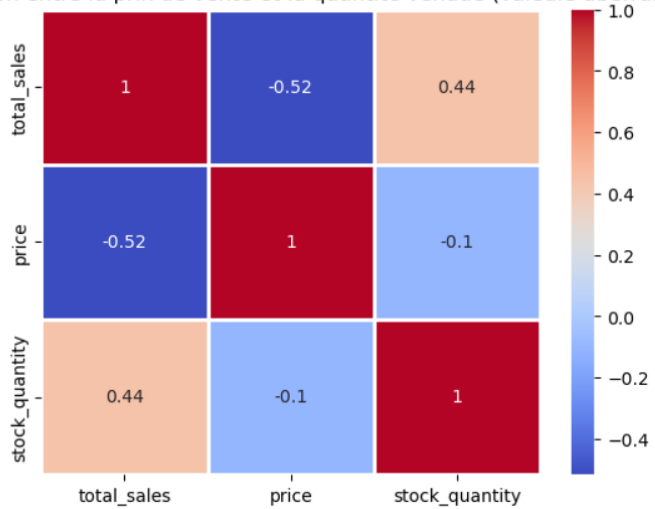




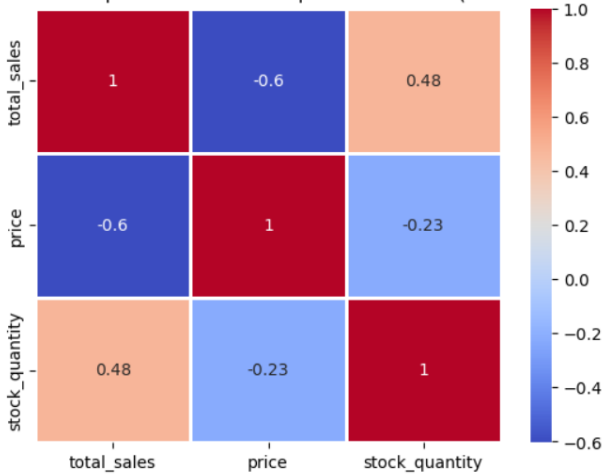


Analyse des corrélations - Heatmaps

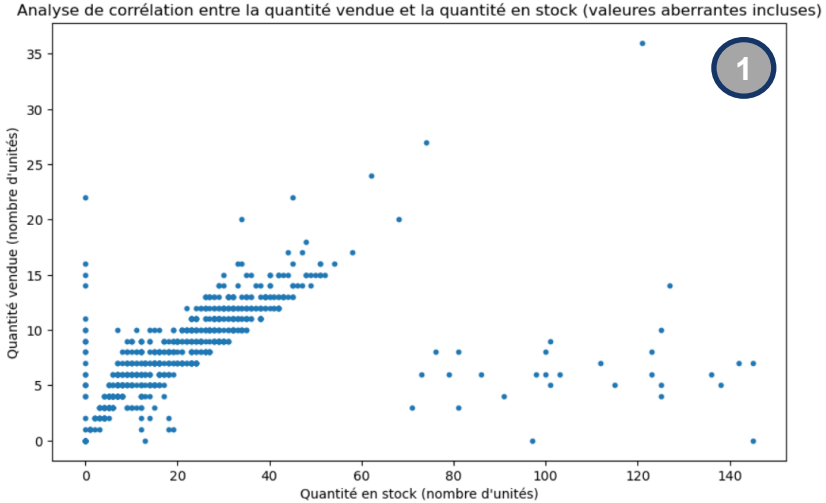
Analyse de corrélation entre la prix de vente et la quantité vendue (valeurs aberrantes incluses)

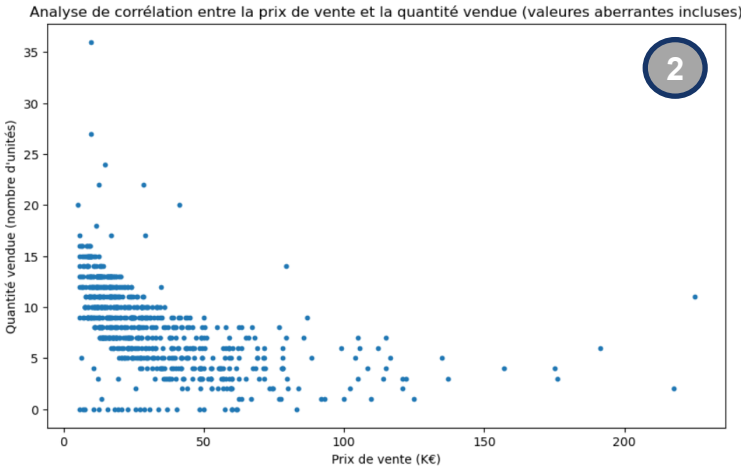


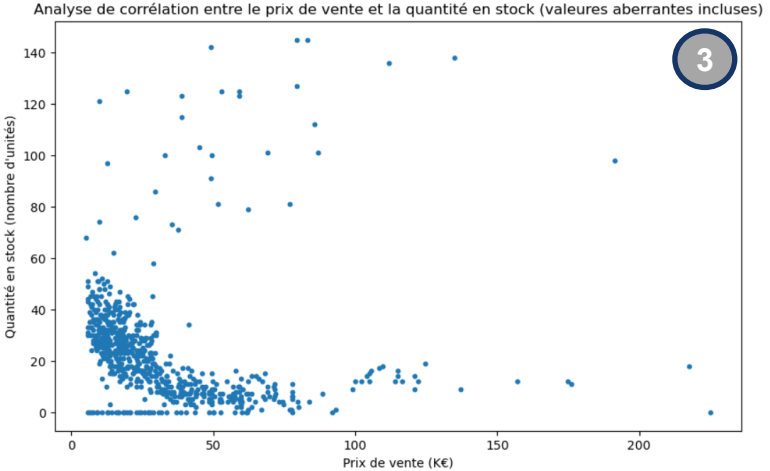
Analyse de corrélation entre la prix de vente et la quantité vendue (valeurs aberrantes exclues)



Analyse des corrélations - Focus **avec** outliers (prix vente)







Des corrélations apparentes

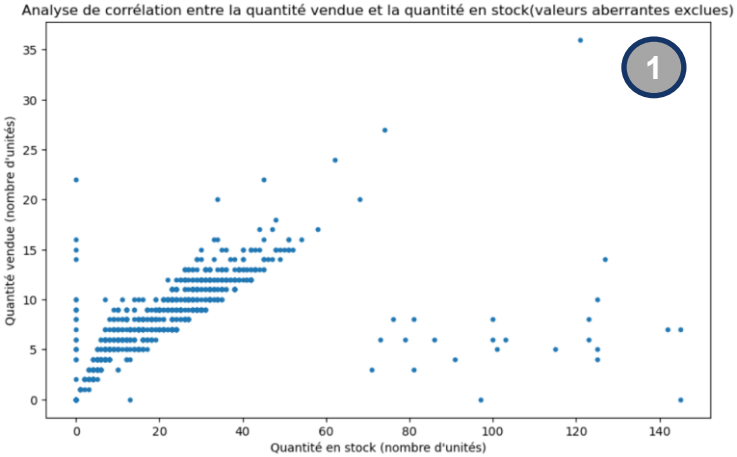
Mais des individus (articles) **volatiles** qui montrent :

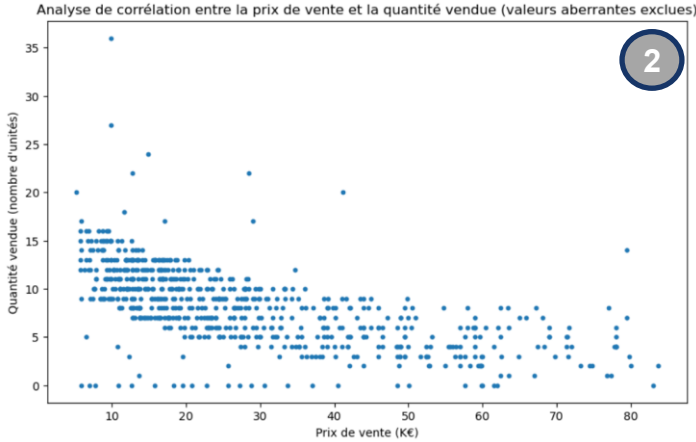
Graphique 1
Des articles en rupture de stock
Des articles avec un stock trop élevé (>60K€)

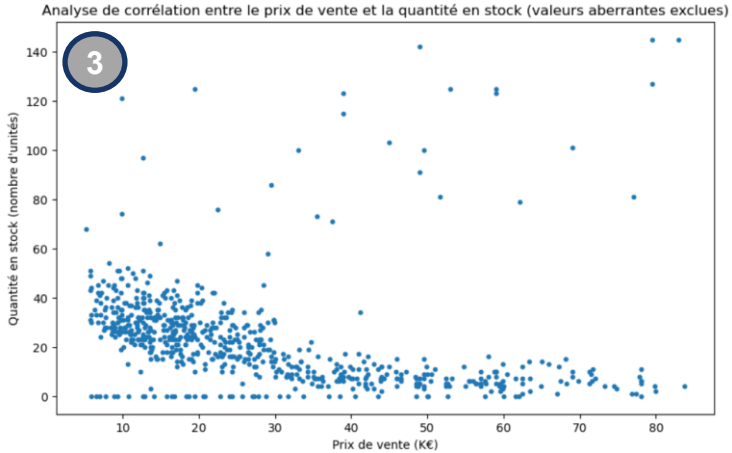
Graphique 2
Des articles non vendus indépendamment des prix de vente
Des articles entre 10 et 40€ vendus avec des quantités sup. à 15 (> 1 carton de 12)

Graphique 3
Des articles en rupture de stock
Des articles avec un stock trop élevé (>60K€) indépendamment du prix de vente

Analyse des corrélations - Focus **sans** outliers (prix vente)







Même observations avec ou sans valeurs aberrantes (outliers)

Mise en cohérence des données dans l'ERP

Corriger les incohérences dans l'ERP ou enrichir les données

- ✓ Sur les identifiants d'articles non renseignés ou non numériques
- ✓ Sur les prix d'achat < 0
- ✓ Sur les prix d'ahcat < 0
- ✓ Quand le prix de vente <= prix d'achat, ou à défaut assumer les marges négatives
- ✓ Sur les stockes < 0
- ✓ Sur les stocks > 0 avec indication d'une rupture de stock
- ✓ Historiser les prix et les stocks par jour pour mise en cohérence avec les dates d'achats quotidiens des articles sur le site
- ✓ Rajouter le taux de TVA applicable à chaque type de produit

Orientations stratégiques

1. **Optimiser la gestion des stocks**
   - ✓ Réduire la durée d'écoulement de certains articles (elle est en moyenne de 40 mois, soit plus de 3 ans)
   - ✓ Réduire (ou supprimer) le stock des articles pour lesquels la durée d'écoulement est de l'ordre de 200 mois, autrement dit 25 ans, autrement dit un stockage infini)
   - ✓ Prévoir un stock pour tous les articles consultés sur le web sans décision d'achat (aucune quantité vendue pour cause de rupture de stock)

2. **Optimiser les ventes**

   - ✓ Elargir et diversifier les références pour les taux de marge les plus élevés
   - ✓ Segmenter la clientèle selon le type de produit : spiritueux, champagnes, vins, etc.
   - ✓ Pour les références à prix élevé, définir pour les clients passionnés, prêts à mettre le prix, un club VIP avec un design et un parcours client adaptés

Limites de l’étude

- ✓ Suppression des lignes avec des biais dans les 3 fichiers, ce qui fausse les analyses
- ✓ Un échantillon représentatif sur 3 ans qui inclut toutes les ventes pour chaque référence (dans notre échantillon une et seule vente par article sur une durée de trois ans)
- ✓ Compte tenu de la limite de l’échantillon qui ne contient qu’une seule vente pour un jour donné, hypothèse d’une augmentation linéaire du CA sur une année pour pour exprimer la durée d’écoulement des stocks en nombre de mois
- ✓ Absence d’un référentiel historisé du catalogue et des stocks (son absence constitue un biais car les stocks et les prix de vente varient avec les temps)
- ✓ Application, pour calculer le taux de marge, d’un taux de TVA de 20% pour convertir le prix de vente TTC en prix de vente HT, alors que toutes les catégories de produit ne sont pas à 20% (Huile d’olive : 5.5%)
- ✓ Comparer les résultats avec une analyse complémentaire sans outliers

ANNEXES

Annexe 2 : analyse du taux de marge y.c. marge négative (un seul article dans l'échantillon)

	Article le moins cher	Article le plus cher	Article le moins rentable	Article le plus rentable
Code article	14570	15940	12589	14774
Libellé article	Moulin de Gassac IGP Pays d'Hérault Guilhem Bl...	Champagne Egly-Ouriet Grand Cru Millésimé 2008	Champagne Egly-Ouriet Grand Cru Blanc de Noirs	Wemyss Malts Single Cask Scotch Whisky Chocola...
Quantité vdue	20	11	0	1
Prix de vente TTC	5.2€	225.0€	12.65€	93.0€
CA	104.0€	2475.0€	0.0€	93.0€
Prix de vente HT	4.33€	187.5€	10.54€	77.5€
Prix d'achat HT	2.74€	137.81€	77.48€	40.49€
Marge	1.59€	49.69€	-66.94€	37.01€
Taux de marge	58.03%	36.06%	-86.4%	91.41%

Point sur les compétences apprises

Le plus facile à appliquer : *nettoyage des données hors détection et correction des données*

Le plus difficile à appliquer :

- ✓ *Neutralisation des doublons, l'argument keep=False de la fonction drop_duplicates n'est pas trivial, il signifie non pas que les doublons sont conservés dans le dataset, mais le contraire. Il eut été plus lisible de positionner keep=True pour conserver les doublons dans le dataset, ou bien d'inverser l'argument en drop=False*
- ✓ *Comprendre et aplliquer pour détecter les valeurs aberrantes la méthode suivante :*
  - ➢ *Visualiser la répartition de la population selon une variable donnée au moyen d'un histogramme pour vérifier si cette répartition suit une loi normale ou non*
  - ➢ *Visualiser la répartition de la population selon une variable donnée au moyen d'une boîte à moustaches pour vérifier si cette répartition contient ou non des valeurs aberrantes*
  - ➢ *Différencier pour identifier les valeurs aberrantes la méthode par la moyenne (z-score) et la méthode par la médiane (quartiles, IQR). Comprendre pourquoi la méthode par le médiane est la plus fiable*
- ✓ *Compréhension de l'utilité et de l'interprétation d'un heatmap pour analyser les corrélations*
- ✓ *Détection des valeurs aberrantes selon les deux méthodes (médiane,*
- ✓ *Appréhender la forme des répartitions pour vérifier mathématiquement si elles suivent ou non une loi normale (skewness, kurtosis)*

Besoin d'approfondissement théorique et d'entraînement : *skewness et kurtosis*