

NETTOYAGE DE LA BDD ET ANALYSE DU PRIX

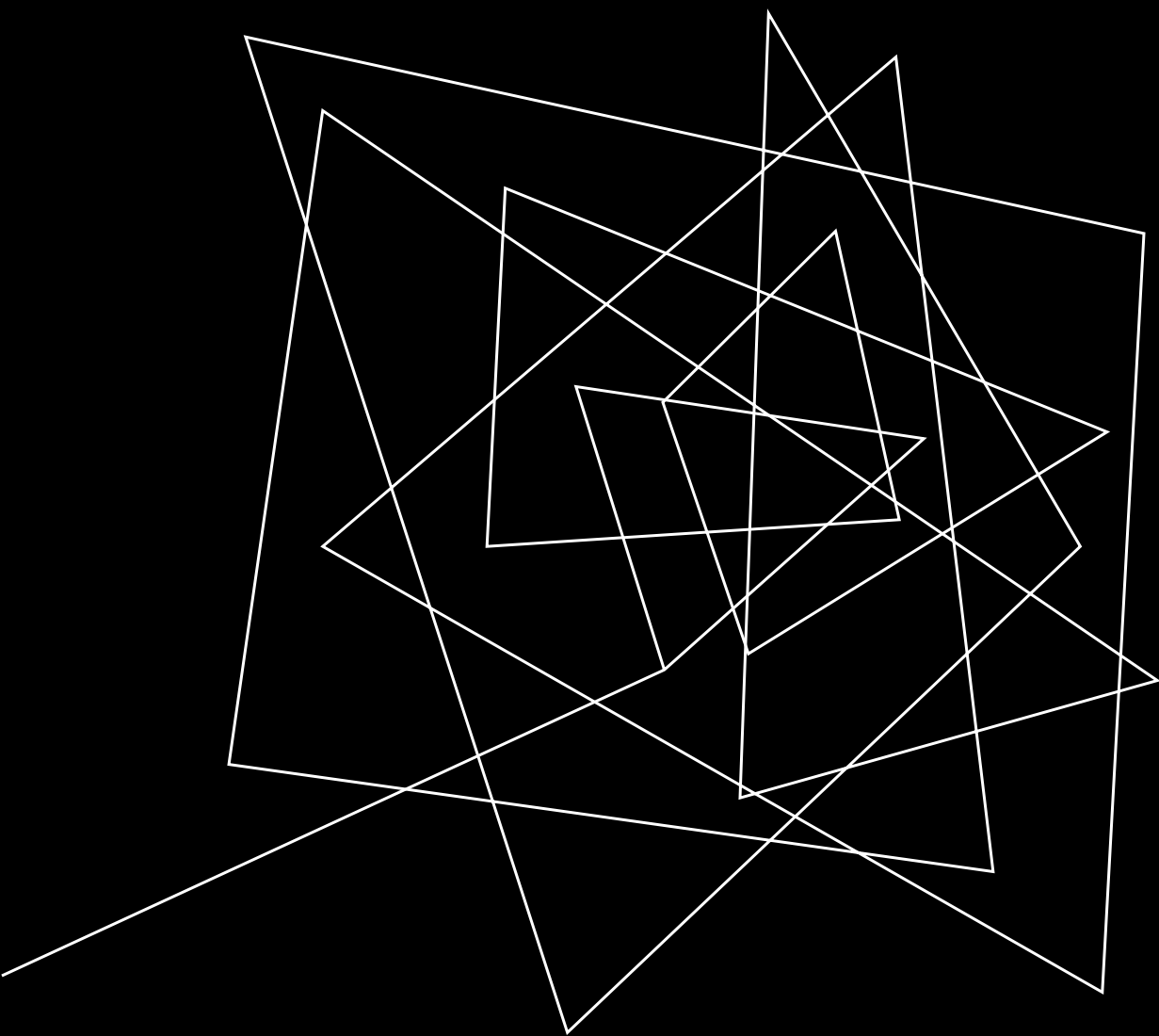
Pauline Felten

CONTEXTE

- L'ERP n'est pas relié au site de vente en ligne => les stocks ne peuvent être gérés efficacement et les ventes sont difficilement analysables.
- La cohérence des données ne peut pas être vérifiée sans liaison entre ces deux jeux de données.
- Possibles erreurs sur certains prix des produits.

METHODOLOGIE DE L'ANALYSE

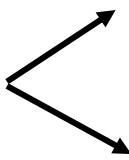
- Rapprocher les jeux de données afin d'avoir toutes les données regroupées en un seul tableau (après contrôle et modifications éventuelles des incohérences identifiées)
- Calculer le CA par produit et le CA réalisé sur le web
- Analyser la distribution de la variable prix pour détecter d'éventuelles erreurs
- Conclusion sur la variable prix



ETAPE 1

Contrôle, nettoyage et modification des
jeux de données

INCOHERENCES RELEVÉES

- Le fichier WEB comporte 1513 lignes pour 714 SKU 
 - 85 lignes dont SKU vides (2 lignes avec des ventes stockées)
 - Lignes doublées pour chaque article soit 714 lignes photos
- Le fichier liaison comporte une erreur au niveau de l'intitulé de colonne SKU
- Un SKU est une chaîne de caractère, un SKU contient un caractère spécial
- Deux valeurs prix sont négatives - aucune autre information
- 91 lignes n'ont pas pu être rapprochées car le SKU est manquant

	Mis en ligne		Statut du stock		Stock (art. en rupture)		
	Oui	Non	En stock	Rupture	Oui	Non	Négatif
Nb de produits	3	88	62	29	4	24	1
Total	91		91		29		

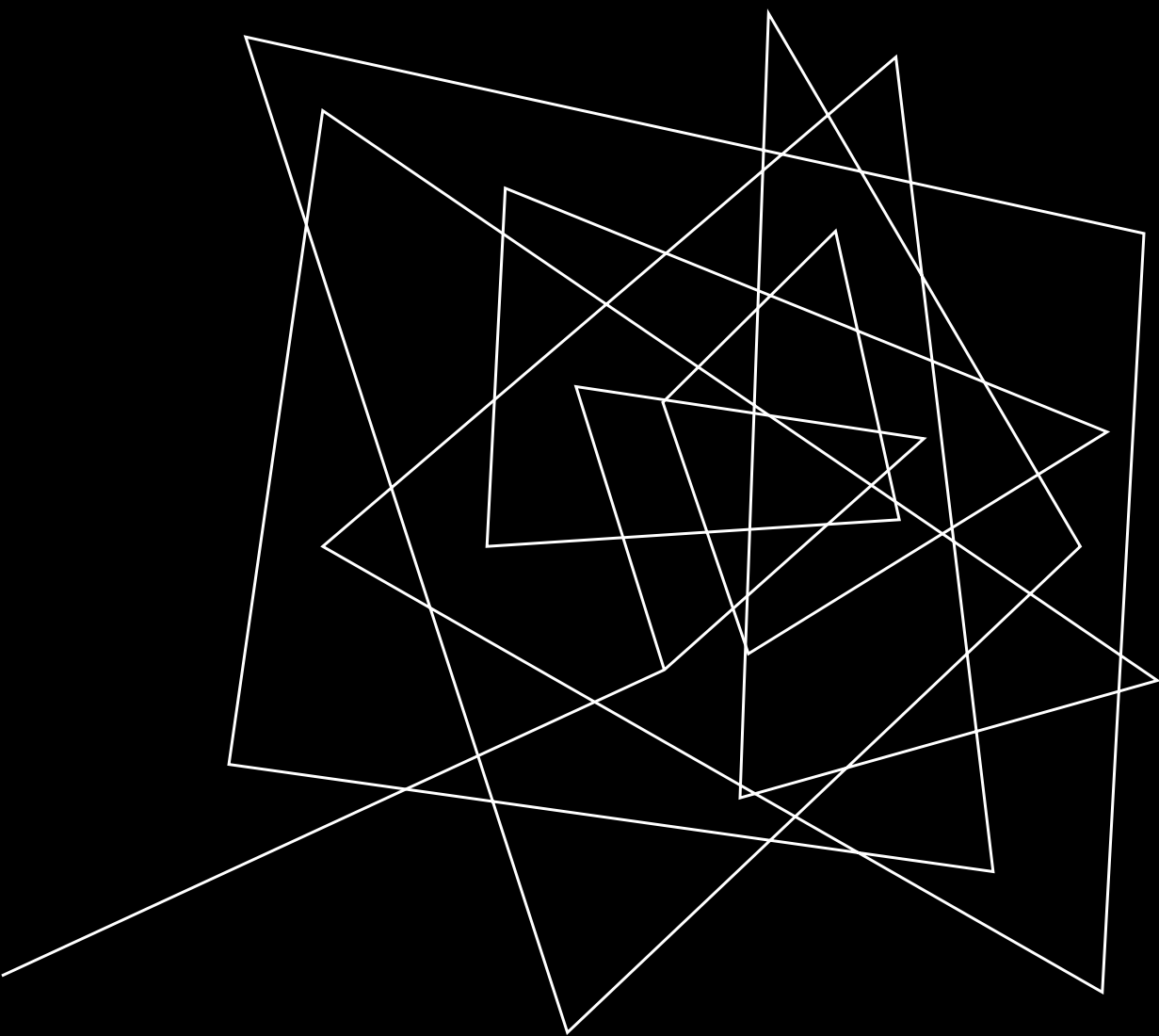
CHOIX DE LA CLÉ PRIMAIRE

	Table de liaison	Fichier ERP	Fichier WEB	
product_id	825	825		
sku	734		714	← 20 SKU manquants
sku manquants	91			

↑
91 product_id n'ont pas de SKU correspondants

Les ID de chaque fichier semblent être un choix pertinent puisque nous disposons d'une table de liaison et que la condition d'unicité pour ces deux clés est vérifiée.

=> Toutefois **product_ID** semble être plus fiable car nous ne perdrons pas de données en l'utilisant.



ETAPE 2

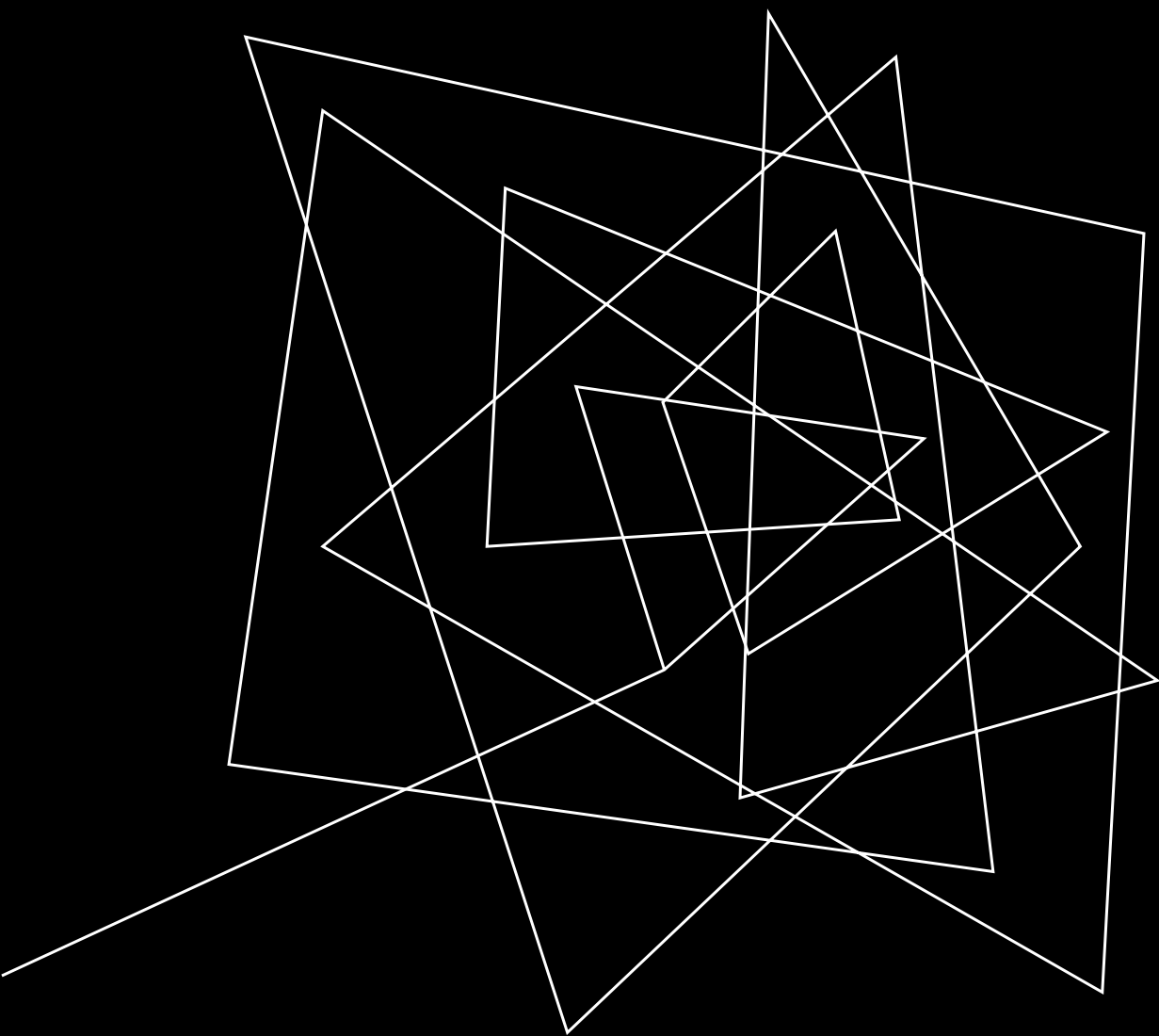
Rapprocher les DataFrames Web et ERP

CHOIX DES JOINTURES

	Table de liaison	Fichier ERP	Fichier WEB	
product_id	825	825		← Inner join
sku	734		714	

Outer join

The diagram illustrates two types of database joins. An inner join is shown between the 'product_id' column and the 'Fichier ERP' column, both containing the value 825. An outer join is shown between the 'sku' column (containing 734) and the 'Fichier WEB' column (containing 714). The 'Fichier WEB' value 714 is bolded, indicating it is a result of the outer join operation.



ETAPE 3

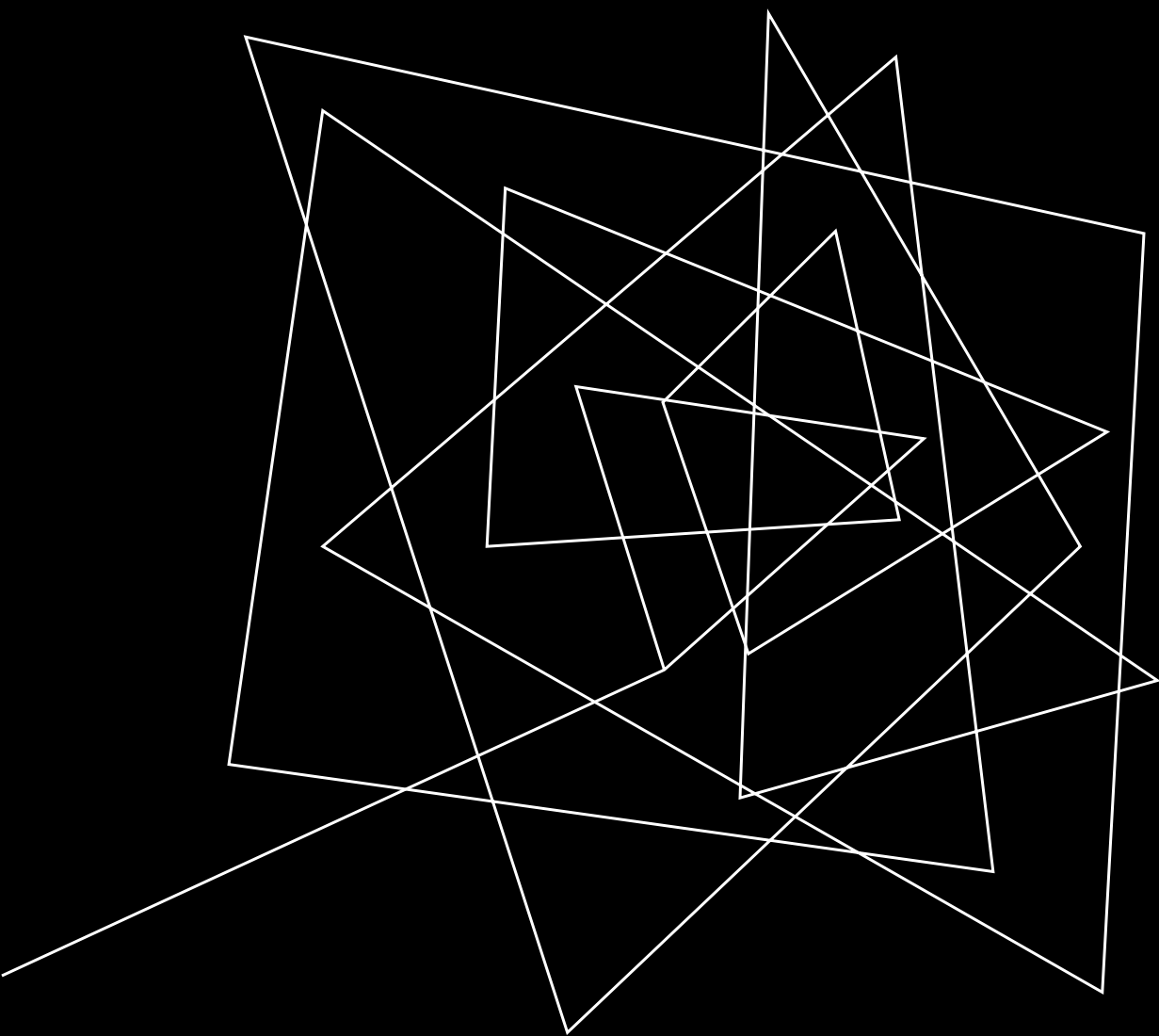
Calcul du CA par produit et CA total
réalisé sur le web

TOP 10
MEILLEURES
VENTES

Product_id	SKU	Nom du produit	Total des ventes	Prix Unitaire	Total CA
4334	7818	Champagne Gosset Grand Blanc de Blancs	96	49,00 €	4 704,00 €
4144	1662	Champagne Gosset Grand Rosé	87	49,00 €	4 263,00 €
4402	3510	Cognac Frapin VIP XO	13	176,00 €	2 288,00 €
4142	11641	Champagne Gosset Grand Millésime 2006	30	53,00 €	1 590,00 €
4141	304	Champagne Gosset Grande Réserve	40	39,00 €	1 560,00 €
4355	12589	Champagne Egly-Ouriet Grand Cru Blanc de Noirs	11	126,50 €	1 391,50 €
4352	15940	Champagne Egly-Ouriet Grand Cru Millésimé 2008	5	225,00 €	1 125,00 €
4153	16237	Elia Daros Côtes du Marmandais Clos Baquey 2015	36	29,00 €	1 044,00 €
6206	16580	Domaine Giudicelli Patrimonio Blanc 2019	41	25,20 €	1 033,20 €
4068	16416	Gilles Robin Crozes - Hermitage Rouge Papillon 2019	62	16,60 €	1 029,20 €



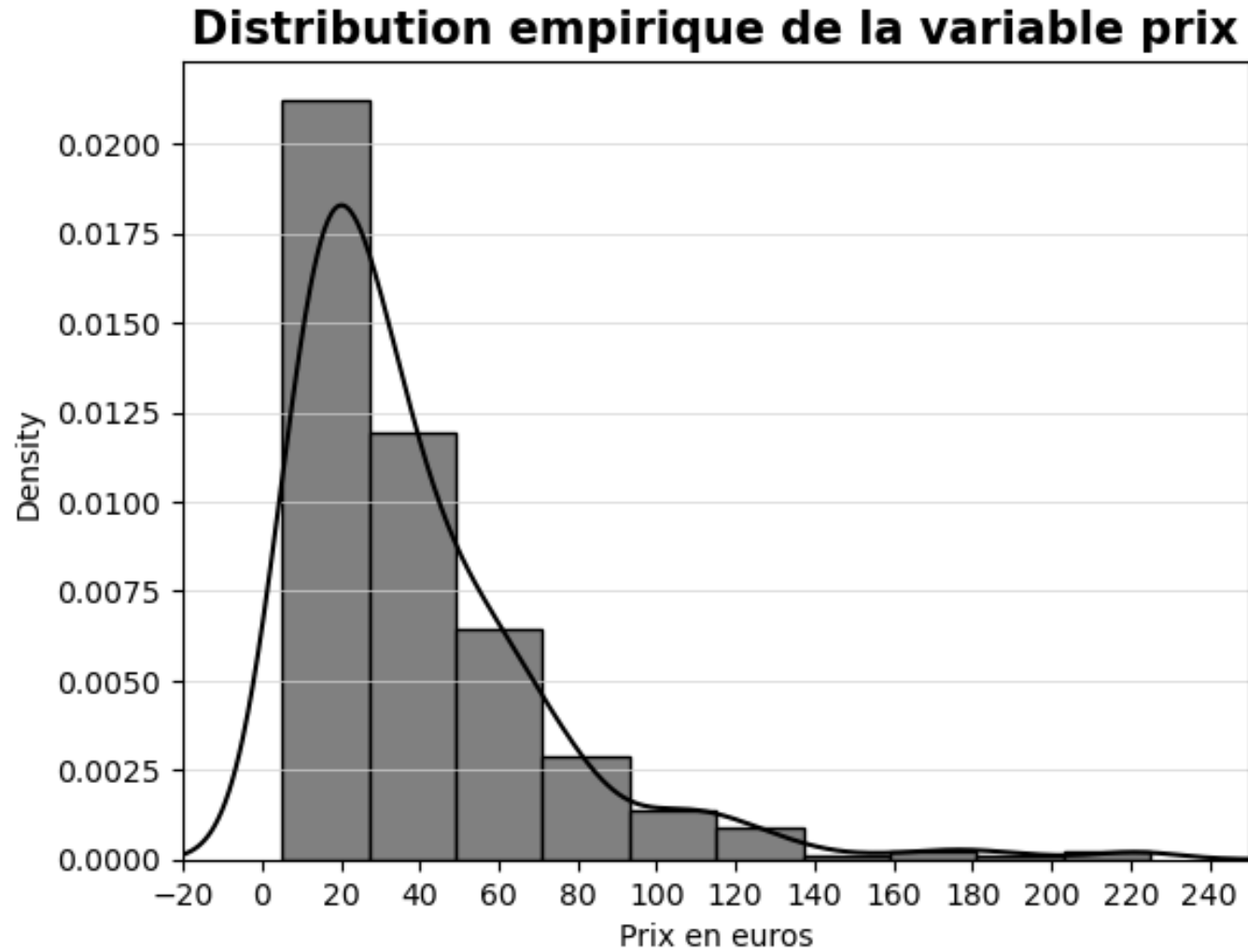
Le CA total réalisé sur le web **70 568,60€**



ETAPE 4

Analyse du prix des produits

DISTRIBUTION EMPIRIQUE



MESURES DE TENDANCE CENTRALE ET MESURES DE DISPERSION

Mode	Moyenne	Médiane
19 €	32.49 €	23.55 €

Etendue	Variance Empirique	Ecart-type	Coef de variation
219.8	772	27.81	85.59%



Les mesures de dispersion montrent que les observations sont dispersées et que le groupe est hétérogène.

MESURES DE FORME

Skewness
2.61



Distribution étalée à droite de la moyenne = La moyenne et la médiane sont plus grand que le mode

Kurtosis
10.56



Distribution leptokurtique = beaucoup d'outliers

OUTLIERS

Deux méthodes utilisées pour la détection d'outliers :

	Ecart interquartile	Z-score
1,5x	32	49
3x	9	14

Concernant les deux valeurs prix négatives, il n'y a aucune information supplémentaire qui nous permet d'en déterminer l'origine. Il s'agit sans doute d'une erreur de frappe.

ECART-INTERQUARTILE

	Prix
Q1	14.10
Q2	23.55
Q3	42.18



Sont considérées comme outliers toutes les valeurs hors des fourchettes ci-dessous

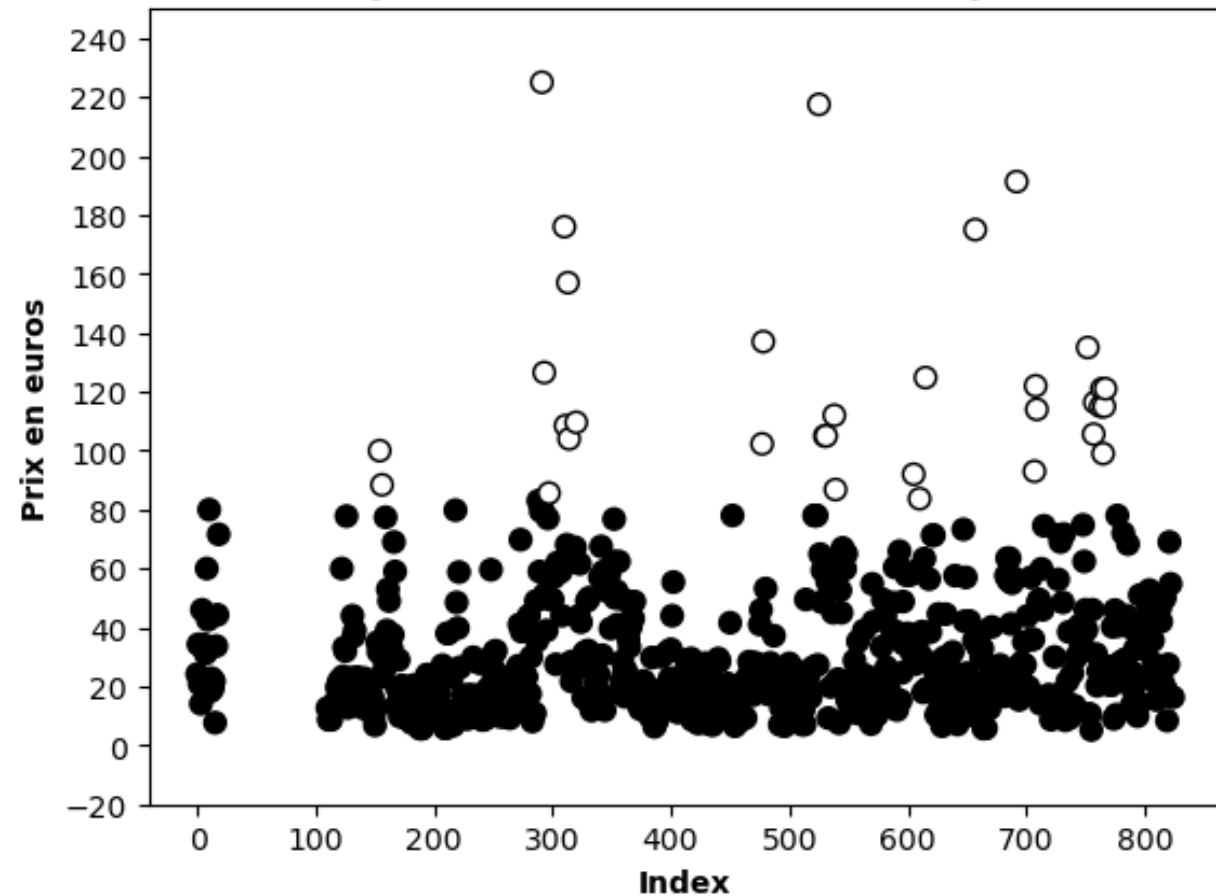
	valeur min	valeur max
limite intérieure (*1,5)	-28	84
limite extérieure (*3)	-70	126

LISTE D'OUTLIERS PAR L'ECART INTERQUARTILE

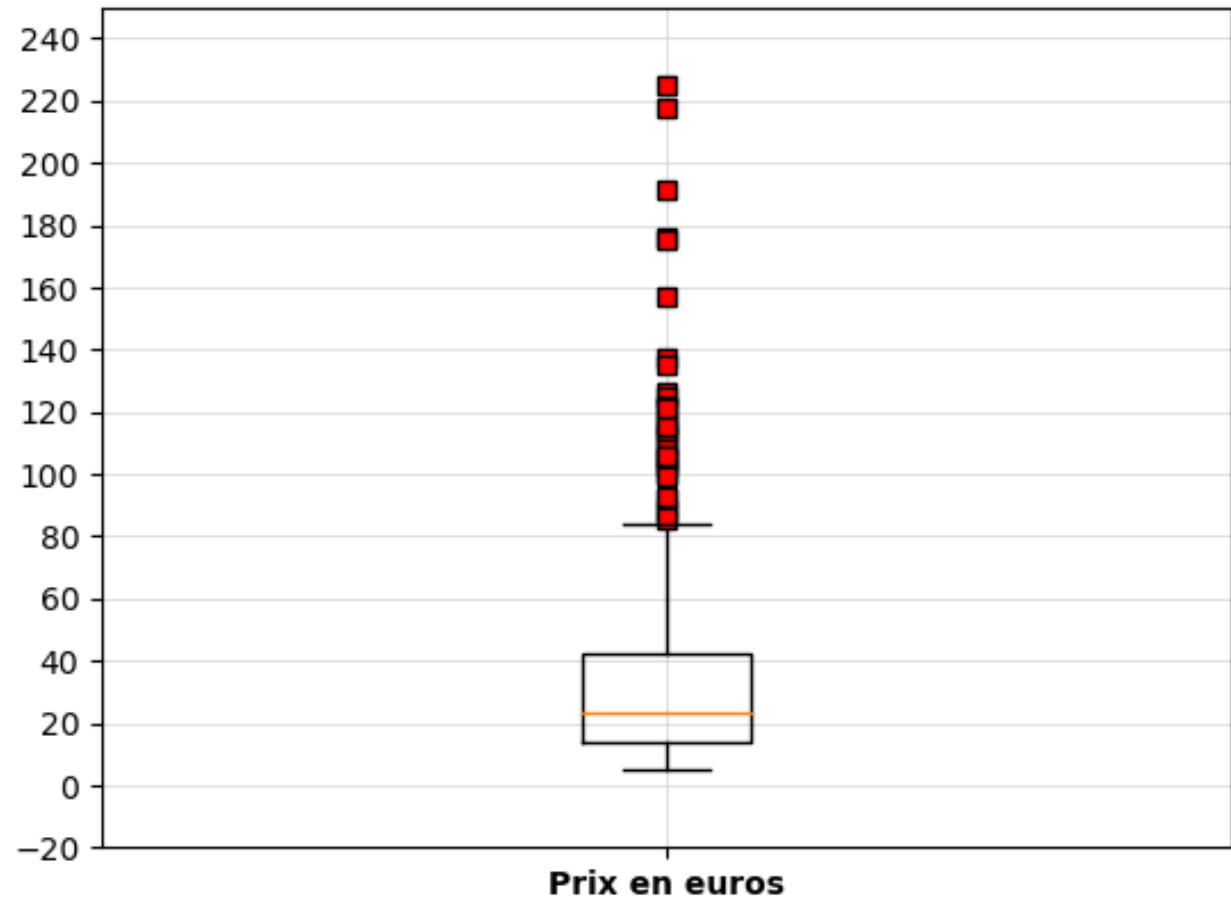
	product_id	sku	post_title	onsale_web	stock_status	stock_quantity	total_sales	price	total_CA
0	4352	15940	Champagne Egly-Ouriet Grand Cru Millésimé 2008	1	outofstock	0	5.0	225.0	1125.0
1	5001	14581	David Duband Charmes-Chambertin Grand Cru 2014	1	instock	20	0.0	217.5	0.0
2	5892	14983	Coteaux Champenois Egly-Ouriet Ambonnay Rouge ...	1	instock	10	3.0	191.3	573.9
3	4402	3510	Cognac Frapin VIP XO	1	instock	8	13.0	176.0	2288.0
4	5767	15185	Camille Giroud Clos de Vougeot 2016	1	instock	12	0.0	175.0	0.0
5	4406	7819	Cognac Frapin Château de Fontpinot 1989 20 Ans...	1	instock	3	0.0	157.0	0.0
6	4904	14220	Domaine Des Croix Corton Charlemagne Grand Cru...	1	instock	13	5.0	137.0	685.0
7	6126	14923	Champagne Gosset Célébris Vintage 2007	1	instock	10	2.0	135.0	270.0
8	4355	12589	Champagne Egly-Ouriet Grand Cru Blanc de Noirs	1	instock	2	11.0	126.5	1391.5

REPARTITION DES OBSERVATIONS DE LA VARIABLE PRIX

Répartition de la variable prix

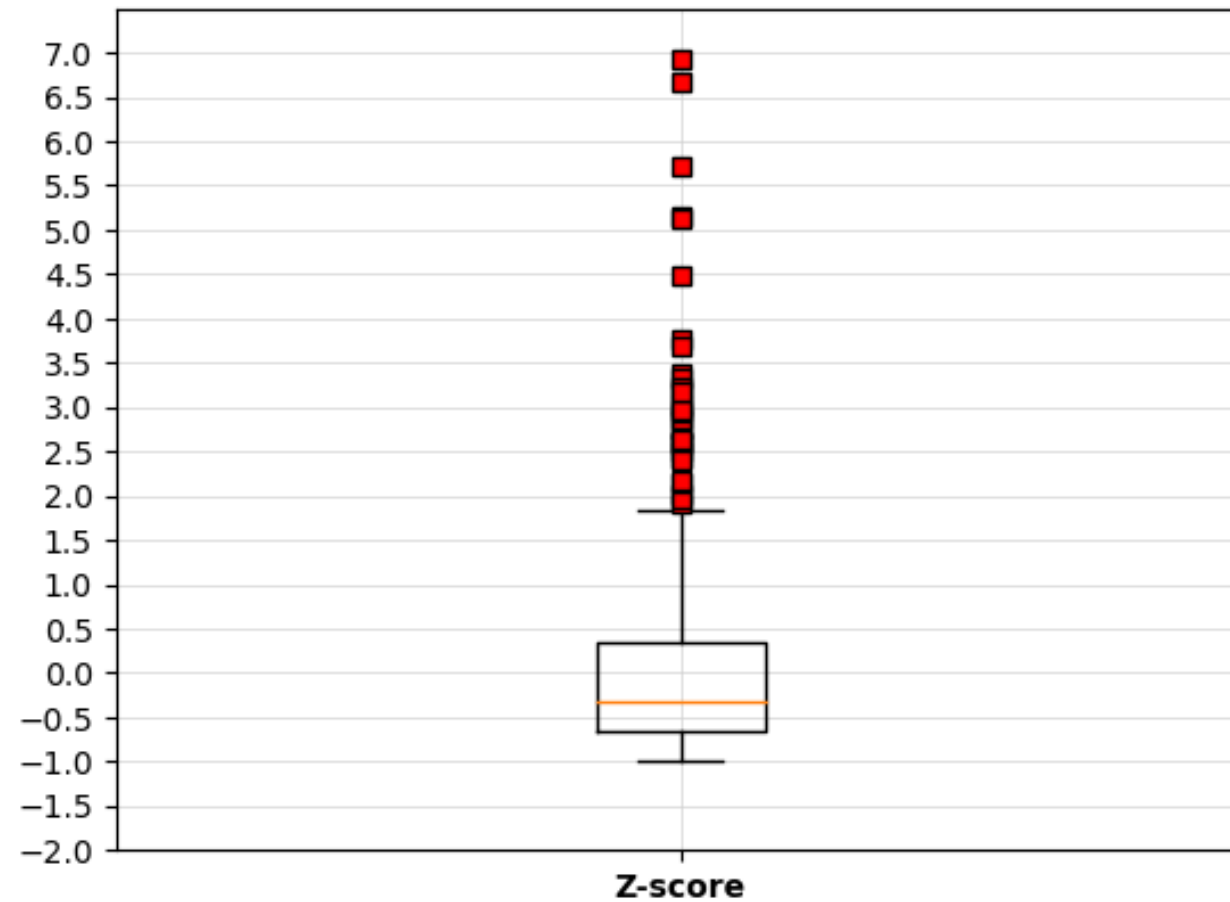


Répartition de la variable prix



Z-SCORE

Répartition de la variable prix selon calcul du Z-score



LISTE D'OUTLIERS PAR LE Z-SCORE

	product_id	sku	post_title	onsale_web	stock_status	stock_quantity	total_sales	price	total_CA
0	4352	15940	Champagne Egly-Ouriet Grand Cru Millésimé 2008	1	outofstock	0	5.0	225.0	1125.0
1	5001	14581	David Duband Charmes-Chambertin Grand Cru 2014	1	instock	20	0.0	217.5	0.0
2	5892	14983	Coteaux Champenois Egly-Ouriet Ambonnay Rouge ...	1	instock	10	3.0	191.3	573.9
3	4402	3510	Cognac Frapin VIP XO	1	instock	8	13.0	176.0	2288.0
4	5767	15185	Camille Giroud Clos de Vougeot 2016	1	instock	12	0.0	175.0	0.0
5	4406	7819	Cognac Frapin Château de Fontpinot 1989 20 Ans...	1	instock	3	0.0	157.0	0.0
6	4904	14220	Domaine Des Croix Corton Charlemagne Grand Cru...	1	instock	13	5.0	137.0	685.0
7	6126	14923	Champagne Gosset Célébris Vintage 2007	1	instock	10	2.0	135.0	270.0
8	4355	12589	Champagne Egly-Ouriet Grand Cru Blanc de Noirs	1	instock	2	11.0	126.5	1391.5
9	5612	14915	Domaine Weinbach Gewurztraminer Grand Cru Furs...	1	instock	12	0.0	124.8	0.0
10	5917	14775	Wemyss Malts Single Cask Scotch Whisky Choc 'n...	1	instock	4	0.0	122.0	0.0
11	6213	15072	Domaine des Comtes Lafon Volnay 1er Cru Santen...	1	instock	7	0.0	121.0	0.0
12	6216	15070	Domaine des Comtes Lafon Volnay 1er Cru Champa...	1	instock	6	0.0	121.0	0.0
13	6202	15126	Domaine Clerget Echezeaux Grand Cru En Orveaux...	1	instock	14	0.0	116.4	0.0

CONCLUSION

Une solution d'harmonisation des données doit être mis en place.

La distribution des prix est étalée car les produits issus du secteur viticole sont présents dans toutes les gammes de prix. Par conséquent, on peut considérer qu'il n'y a pas de valeurs aberrantes dans nos données.