

Projet 6

Optimisez la gestion des données
d'une boutique avec R ou Python

Boutique de vins Bottleneck



Marie G.
Parcours Data Analyst
19/02/2025

1. Analyse exploratoire des données



1.1. Fichiers de données analysés

Données Bottleneck récoltées

Fichier excel  →  Dataframe pandas

Extraction de l'ERP

erp.xlsx → df_erp

Extraction du site web

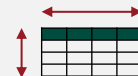
web.xlsx → df_web

Table de liaison entre id des 2 tables

liaison.xlsx → df_liaison

Analyse exploratoire menée pour chaque fichier

- Analyse des valeurs renseignées (colonnes) et du nombre d'observations (lignes)
- Vérification des types attendus de valeurs dans chaque colonne
- Recherche de doublons
- Recherche de valeurs nulles
- Vérification de la cohérence des valeurs



5	XX	2	True
2	XX	1	0
abc	XX	10.1	False

A	C	T	J
XX	XX	XX	XX
A	C	T	J

XX	XX	XX	XX
XX	NaN	XX	XX
XX	XX	XX	XX

- Le **nettoyage** présenté ici est proposé en première approche et **devra être validé par les équipes opérationnelles et la DSI**
- Il conviendra **d'identifier avec les équipes d'où proviennent les incohérences observées** sur le site internet et dans l'ERP

1. Analyse exploratoire des données



1.2. Analyse exploratoire de l'extraction de l'ERP *erp.xlsx* (1/2)

1 - Caractéristiques générales de *df_erp*

Column	product_id	onsale_web	price	stock_quantity	stock_status	purchase_price
Non-Null	825	825	825	825	825	825
Count	non-null	non-null	non-null	non-null	non-null	non-null
Dtype	int64	int64	float64	int64	object	float64

Analyses :

- 825 valeurs
- Aucune valeur nulle
- Pas d'erreurs de type
- **Identifiants '*product_id*' uniques : clé de la table**

2 – Analyse des la variable '*stock_status*'

- Incohérences entre la valeur de '*stock_quantity*' et celle de '*stock_status*' :

	product_id	onsale_web	price	purchase_price	stock_quantity	stock_status
4	4039	1	46.0	23.77	3	outofstock
398	4885	1	18.7	9.66	0	instock
449	4973	0	10.0	4.96	-10	outofstock
573	5700	1	44.5	22.30	-1	outofstock

Nettoyage réalisé :

- '*stock_quantity*' = 0 ➡ '*stock_status*' = 'outofstock'
- '*stock_quantity*' < 0 ➡ '*stock_quantity*' = 0 & '*stock_status*' = 'outofstock'
- '*stock_quantity*' > 0 ➡ '*stock_status*' = 'instock'

➔ **Dysfonctionnement à identifier dans l'ERP**

1. Analyse exploratoire des données



1.2. Analyse exploratoire de l'extraction de l'ERP *erp.xlsx* (2/2)

4 – Analyse de la variable 'price'

- Valeur minimum positive : 5,20 €
- Valeur maximum : 225 €
- Identification des prix négatifs :

	product_id	onsale_web	price	purchase_price	stock_quantity	stock_status
151	4233	0	-20.0	10.33	0	outofstock
469	5017	0	-8.0	4.34	0	outofstock
739	6594	0	-9.1	4.61	19	instock

Nettoyage réalisé :

- 'price' < 0 ➡ 'price' = valeur absolue ('price')

➔ Erreur de saisie qui devrait lever une erreur

5 – Analyse de la variable 'stock_quantity'

- Valeur minimale : 0
- Valeur maximale : 145
- Valeur négatives rectifiées lors de l'analyse de la variable 'stock_status'

6 – Analyse de la variable 'onsale_web'

Valeurs de onsale_web :

- '1' : produits vendus en ligne (716 produits)
- '0' : produits non vendus en ligne (109 produits)

7 – Analyse de la variable 'purchase_price'

- Pas de prix d'achat non renseignés
- Pas de prix d'achat négatifs
- Prix d'achat minimum : 2,74 €
- Prix d'achat maximum : 137,81 €

Bilan des erreurs rencontrées et du nettoyage réalisés sur le fichier ERP :

- Passage en valeur absolue des prix de vente négatifs (3 valeurs)
- Remplacement par 0 des stocks négatifs (2 valeurs)
- Rectification de la colonne 'stock_status' en fonction de 'stock_quantity' (2 valeurs)

1. Analyse exploratoire des données



1.3. Analyse exploratoire de l'extraction du site internet web.xlsx (1/3)

1 – Caractéristiques générales du dataset et analyse des colonnes

	Nom de colonne	Non-Null	Dtype	Valeurs
0	sku	1428	object	Identifiant web du produit
5	total_sales	1430	float64	Nombre de ventes en ligne
8	post_author	1430	float64	Créateur du post : '1.0' (2) / '2.0' (1428)
9	post_date	1430	datetime64[ns]	Date de mise en ligne
12	product_type	1429	object	Type de produit : vin, champagne, whisky, ...
13	post_title	1430	object	Nom du produit
14	post_excerpt	716	object	Descriptif du produit
20	post_modified	1430	datetime64[ns]	Date de dernière modification du produit
24	guid	1430	object	Selon 'post_type', lien url vers la page du produit ou vers une image
26	post_type	1430	object	Type de lien de la colonne 'guid' : 'product' (716) 'attachment' (714)

Premières analyses :

- 29 colonnes dont 10 conservées
- 1513 lignes dont 1430 non nulles conservées
- 714 valeurs distinctes de sku : **clé identifiante du produit web**
- La majorité des produits sont renseignés en doublon (voir 'sku' ci-dessous)**
 - Valeur 'product' dans 'post_type' -> lien vers la page web produit donné dans 'guid'
 - Valeur 'attachment' dans 'post_type' -> lien vers l'image du produit donné dans 'guid'

2 – Analyse de la variable 'sku' / 'id_web' (clé de la table)

- Valeurs entières comprises entre 38 et 19822
- 714 doublons pour 1430 lignes dans la table
- 'id_web' non renseignés (seuls produits n'apparaissant pas en doublon) :

	id_web	total_sal es	post_aut hor	post_date	product_type	post_title
1493	NaN	-56.0	2.0	08/08/2018 11:23:43	Vin	Pierre Jean Villa Condrieu Jardin Suspendu 2018
1495	NaN	-17.0	2.0	31/07/2018 12:07:23	Vin	Pierre Jean Villa Côte Rôtie Fongeat 2017

- 'id_web' non conformes :

	id_web	total_sales	post_author	post_date	product_type	post_title
0	bon-cadeau-25-euros	7.0	1.0	01/06/2018 13:53:46	Autre	Bon cadeau de 25€
2	13127-1	4.0	2.0	09/06/2020 15:42:04	Vin	Clos du Mont-Olivet Châteauneuf-du-Pape 2007

Nettoyage réalisé :

- Affectation de nouveaux 'id_web' à ces 4 produits (19823 à 19825)

1. Analyse exploratoire des données



1.3. Analyse exploratoire de l'extraction du site internet web.xlsx (2/3)

3 – Analyse de la variable 'sku' / 'id_web' : suppression des doublons

- Utilisation d'un pivot pour enregistrer sur une même ligne toutes les données ayant le même id_web et pouvoir les comparer
- Quelques données diffèrent entre les lignes en doublon :**
 - 710 où seules les colonnes 'guid' et 'post_type' présentent des différences (exemples)

id_web	guid		post_type	
	0	1	0	1
38	https://www.bottle-neck.fr/?post_type=product&p=4729	https://www.bottle-neck.fr/wp-content/uploads/2020/03/emile-boeckel-cremant-brut-blanc-de-blancs.jpg	product	attachment
41	https://www.bottle-neck.fr/?post_type=product&p=4634	https://www.bottle-neck.fr/wp-content/uploads/2020/03/marcel-windholtz-eau-de-marc-de-gewurztraminer.jpg	product	attachment

Analyse :

- Les produits sont systématiquement renseignés en doublon sur le site web** pour pouvoir renseigner dans le champ 'guid', d'une part un lien vers la page web du produit, de l'autre une image du produit
- ➔ **Il conviendrait de modifier le site web pour pouvoir intégrer en un seul enregistrement ces deux informations**

Nettoyage réalisé :

- On crée une nouvelle table consolidée qui ne conserve qu'un exemplaire de chaque doublon (car identiques)
- On retient en revanche les 2 versions de la valeur 'guid' dans la table consolidée et on supprime 'post_type'

- 4 lignes où, de plus, la colonne 'total_sales' présente des différences

id_web	total_sales		post_author	post_date	product_type	post_title
	0	1				
1366	6.0	116.0	2.0	13/02/2018 13:45	Champagne	Champagne Mailly Grand Cru Intemporelle 2010
14561	11.0	111.0	2.0	01/09/2018 15:34	Vin	Argentine Mendoza Alamos Torrontes 2017
14950	22.0	122.0	2.0	18/04/2018 11:53	Vin	François Baur Pinot Noir Schlittweg 2017
15346	2.0	22.0	2.0	31/07/2018 11:49	Vin	Albert Mann Pinot Noir Grand H 2017

Nettoyage réalisé :

- On retient pour 'total_sales' la colonne 1, dont la colonne 0 semble être une version tronquée
- L'origine de l'erreur serait à analyser avec le gestionnaire du site web

1. Analyse exploratoire des données



1.3. Analyse exploratoire de l'extraction du site internet web.xlsx (3/3)

4 – Analyse de la variable 'total_sales'

- Valeur maximale : 122
- Valeur minimale positive : 0
- 2 valeurs négatives :

id_web	total_sales	product_type	post_title	product_id_from_guid	id_web
19825	-56.0	Vin	Pierre Jean Villa Condrieu Jardin Suspendu 2018	5075	19825
19826	-17.0	Vin	Pierre Jean Villa Côte Rôtie Fongean 2017	5070	19826

Analyse :

- Il s'agit des lignes qui n'avaient originellement pas d'id_web. On suppose que la saisie de ces produits n'a pas suivi le process habituel: il peut s'agir d'une erreur de saisie

Nettoyage réalisé :

- On suppose une erreur de saisie -> **passage en valeur absolue du nombre de ventes**

5 – Analyse de la variable 'post_title'

3 noms de produits apparaissent en double :

- Clos du Mont-Olivet Châteauneuf-du-Pape 2007
- Domaine Hauvette IGP Alpilles Jaspe 2017
- Marc Colin Et Fils Chassagne-Montrachet Blanc Les Vide-Bourses 1er Cru 2016
- ➔ Conservés ici, mais il conviendrait de vérifier si ces articles sont réellement différents

6 – Analyse des variables 'post_date' et 'post_modified'

- Toutes les dates sont renseignées
- 'post_date' (date de publication d'un produit) :
 - Valeur minimale : 2018-02-08
 - Valeur maximale : 2020-07-20
- 'post_modified' (dernière modification) :
 - Valeur minimale : 2018-02-20 15:19:23
 - Valeur maximale : 2020-08-27 18:55:03

7 – Analyse de la variable 'product_type'

- Valeurs prises par la variable :

Vin	660
Champagne	28
Whisky	14
Cognac	8
Huile d'olive	3
Cin	2
Autre	1

8 – Analyse de la variable 'product_id_from_guid'

- Colonne créée à partir de l'url du produit de la colonne 'guid'
(exemple : https://www.bottle-neck.fr/?post_type=product&p=4729)
- Valeur minimale : 3847
- Valeur maximale : 7338
- **Tous les product_id sont renseignés**

1. Analyse exploratoire des données



1.3. Analyse exploratoire de la table de liaison liaison.x/sx

- 825 lignes et 2 colonnes : product_id et id_web
- 825 valeurs de product_id, toutes uniques, qui correspondent exactement aux 825 valeurs de df_erp**
- 734 valeurs de id_web**, toutes uniques, soit 18 de plus que df_web : produits qui ne sont plus vendus en ligne, ou erreur de saisie ?
- 2 id_web ont été modifiés dans df_web ('bon-cadeau-25-euros' et '13127-1') : on les modifie de la même manière dans df_liaison

	id_web	product_id
0	19823	4954
2	19824	7247

- 2 id_web ont été créés dans df_web : on retrouve le product_id grâce à la colonne 'product_id_from_guid' et on modifie ainsi df_liaison :

	id_web	product_id
762	19826	5070
763	19825	5075

- Après ces modifications, on retrouve bien tous les produits de df_web et de df_erp dans la table de liaison. Cependant, **il reste 20 valeurs de la table de liaison qu'on ne retrouve pas dans df_web** :

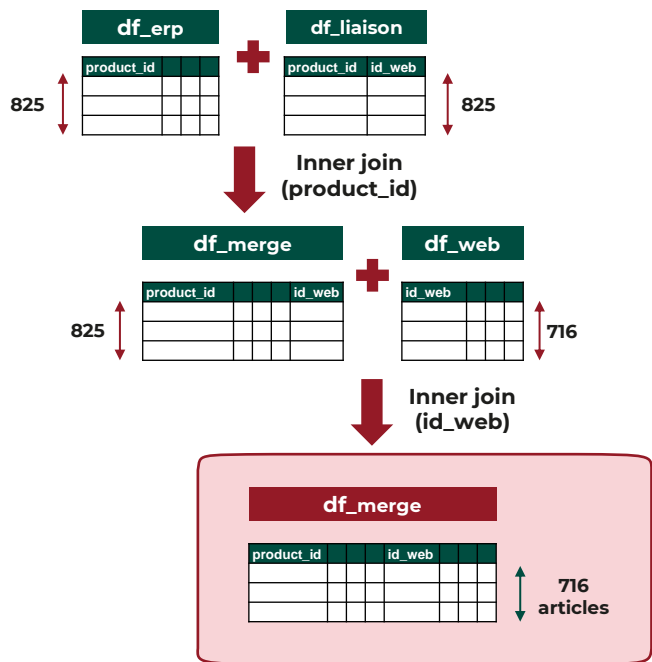
	id_web	product_id
1	14680-1	7329
295	15630	5021
302	15609	5954
303	15608	4921
305	15586	4922
321	15529	6100
401	15272	5018
430	15154	4864
453	15065	4568
497	14785	4584
506	14730	5570
509	14715	5559
515	14689	5800
521	14648	5505
547	14379	5953
548	14377	5955
553	14360	4869
594	13771	4289
606	13577	5957
652	12601	4741

Sans plus d'informations, on conserve ces données dans la table

2. Fusion et consolidations des données



Création d'une table df_merge comportant les données de vente du site web et les données de l'ERP : fusion des données par 2 jointures internes successives, pour obtenir la table df_merge contenant les uniquement données des ventes des produits vendus en ligne :



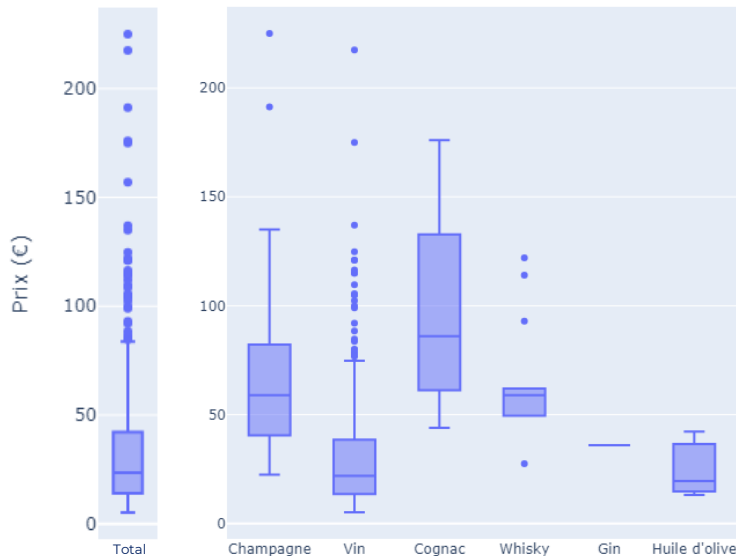
Consolidation de la table df_merge

- Vérification que :
 - tous les articles de df_web se trouvent dans df_merge
 - tous les product_id et id_web sont renseignés
 - la valeur 'product_id' est bien égal au 'product_id_from_guid' pour tous les articles
- Rectification des erreurs sur la valeur 'onsale_web' dans df_erp - 2 produits concernés (product_id) :
 - 4594 ➡ 'onsale_web' = 0
 - 4200 ➡ 'onsale_web' = 1
- Suppression du bon cadeau qui n'est pas pertinent pour les analyses de ventes (715 articles restants)
- Définition d'un index : product_id

3. Analyse univariée du prix

- Moyenne des prix : 32 €
- Ecart-type : 28 €
- Seuil prix dont z-score ≥ 3 : 116,40 € (13 articles)

Prix des articles : boîte à moustache



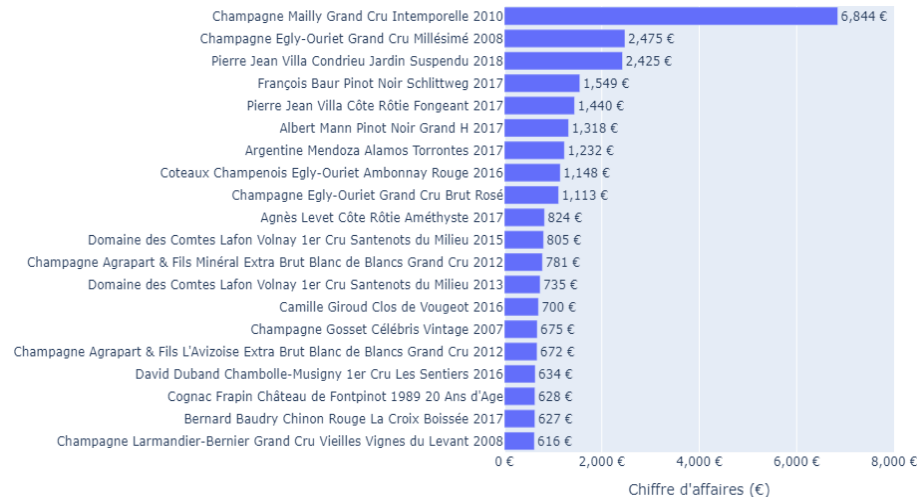
- **13 articles pouvant être considérés comme outliers** (z_score ≥ 3 soit prix ≥ 113 €)
- **Cependant, leur prix est cohérent avec le prix d'alcools d'une certaine réputation (champagnes, vins grands crus, cognac de 20 ans d'âge...)**
- Une comparaison avec des produits similaires vendus en ligne permettrait de lever le doute sur la cohérence de ces prix

product_id	post_title	product_type	purchase_price	price	total_sales	stock_quantity
4352	Champagne Egly-Ouriat Grand Cru Millésimé 2008	Champagne	138 €	225 €	11	0
5892	Coteaux Champenois Egly-Ouriat Ambonnay Rouge 2016	Champagne	116 €	191 €	6	98
5767	Camille Giroud Clos de Vougeot 2016	Vin	90 €	175 €	4	12
6126	Champagne Gosset Célébris Vintage 2007	Champagne	80 €	135 €	5	138
4406	Cognac Frapin Château de Fontpinot 1989 20 Ans d'Age	Cognac	69 €	157 €	4	12
6202	Domaine Clerget Echezeaux Grand Cru En Orveaux 2015	Vin	63 €	116 €	5	12
4402	Cognac Frapin VIP XO	Cognac	78 €	176 €	3	11
5001	David Duband Charmes-Chambertin Grand Cru 2014	Vin	117 €	218 €	2	18
4904	Domaine Des Croix Corton Charlemagne Grand Cru 2016	Vin	68 €	137 €	3	9
5917	Wemyss Malts Single Cask Scotch Whisky Choc 'n' Nut Pretzel 2001 Bunnahabhain	Whisky	54 €	122 €	3	12
6213	Domaine des Comtes Lafon Volnay 1er Cru Santenots du Milieu 2016	Vin	63 €	121 €	3	9
6216	Domaine des Comtes Lafon Volnay 1er Cru Champans 2016	Vin	60 €	121 €	2	14
5612	Domaine Weinbach Gewurztraminer Grand Cru Furstentum SGN 2010 1/2	Vin	66 €	125 €	1	19

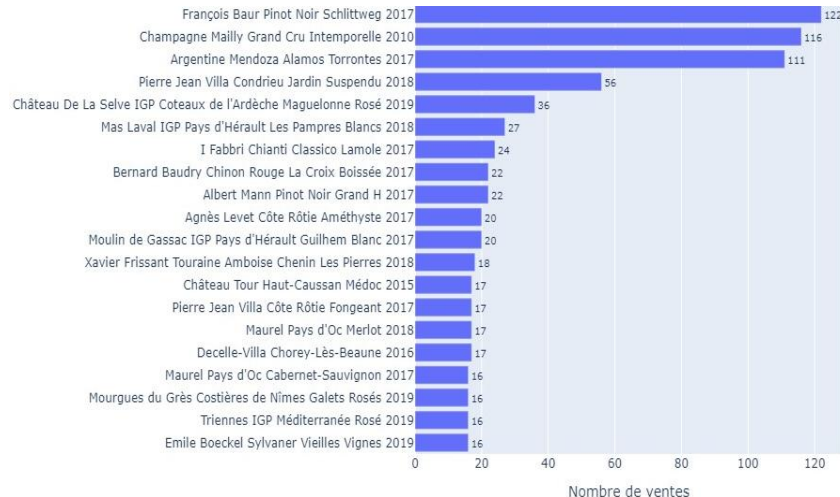
4. Analyse univariée du chiffre d'affaires et des ventes



Chiffre d'affaires par produit du top 20 (octobre 2020)



Nombre de ventes par produit du top 20 (octobre 2020)



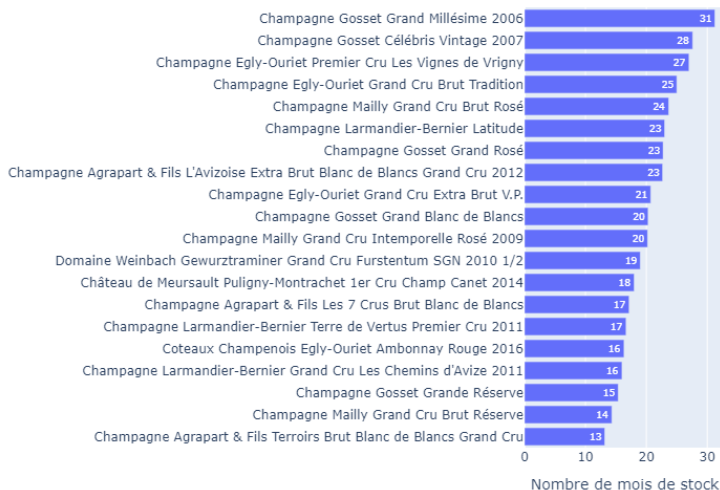
- Chiffre d'affaires total du mois d'octobre 2020 : 157 438 euros
- 58% des articles (soit 417) représentent 80% du chiffre d'affaires du site web

- 59% des articles (soit 423) représentent 80% du chiffre d'affaires du site web

5. Analyse univariée du stock et de la marge



Nombre de mois de stock des produits du flop 20 (octobre 2020)

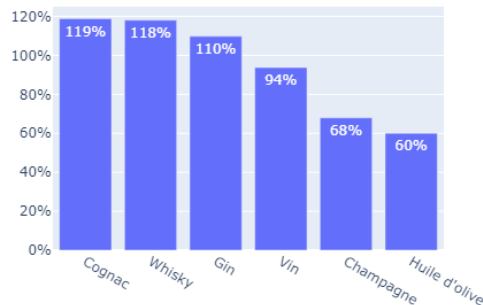


Au 31 octobre 2020, l'état des stocks est le suivant :

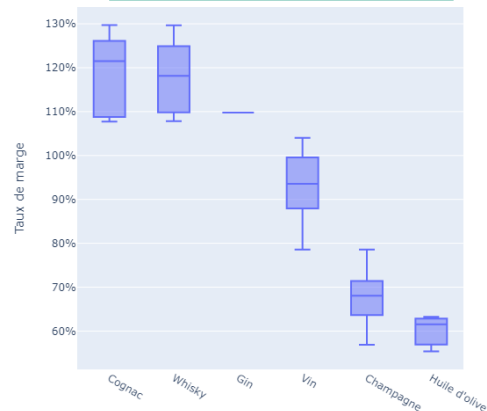
- Valeur des stocks : 494 063 €
- 16 717 produits en stock

Nota : les produits non vendus ne sont pas représentés car le nombre de mois de stock ne peut pas être calculé

Marge moyenne par type de produit



Marge par type de produit : boîte à moustache



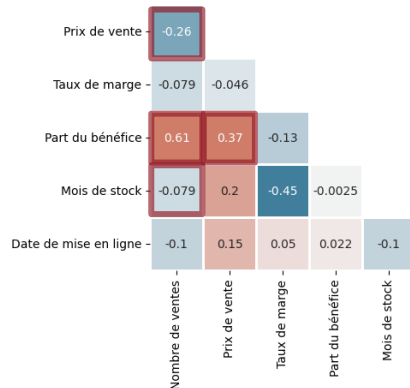
- Taux de marge = $\frac{\text{prix de vente HT} - \text{prix d'achat (HT)}}{\text{prix d'achat}}$
 - Taux de marge positif minimum : 55%
 - Taux de marge maximum : 130%
 - Marge totale octobre 2020 : 62 223 euros
 - Un produit de marge négative (-84%) : erreur de saisie ?
- ➔ Outlier supprimé dans l'analyse multivariée (714 produits restants)

product_id	post_title	Prix HT	Prix d'achat (HT)	Taux de marge	stock_quantity	Valorisation des stocks	total_sales	post_date	post_modified
4355	Champagne Egly-Ouriel Grand Cru Blanc de Noirs	10,54 €	77,48 €	-84%	97	1 227 €	0	02/03/2018 10:46:10	13/08/2020 10:15:02

6. Analyse multivariée

6.1 Analyse par le coefficient de corrélation linéaire de Pearson

Heatmap de corrélation entre variables caractéristiques de chaque article



Déterminer les corrélations entre variables caractéristiques des articles vendus

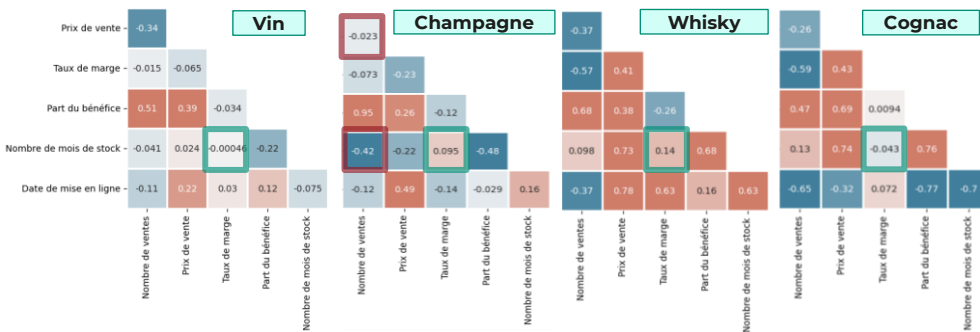
- **Consolider la stratégie de rotation des stocks**
- **Déterminer les caractéristiques des produits participant à maximiser le bénéfice**

Suppression préalable des 22 articles non vendus qui ne sont pas en stock, qui brouilleraient les résultats

- Corrélation positive marquée entre part du bénéfice et nombre de ventes (0.61), et prix de vente (0.37) : **le nombre de ventes impacte plus le bénéfice qu'un prix élevé**
- Corrélation négative marquée entre prix et nombre de ventes (-0.26) : **les produits les plus chers sont les moins vendus - sauf pour le champagne (-0.023)**
- Faible corrélation entre:
 - Prix et marge (-0.045) : **on marge autant sur les produits chers que sur les produits bon marché**
 - Mois de stock et prix de vente (-0.079) : **bonne gestion des stocks a priori sauf pour le champagne (-0.42)**

→ **Corrélations à confirmer grâce à une visualisation par des diagrammes à points**

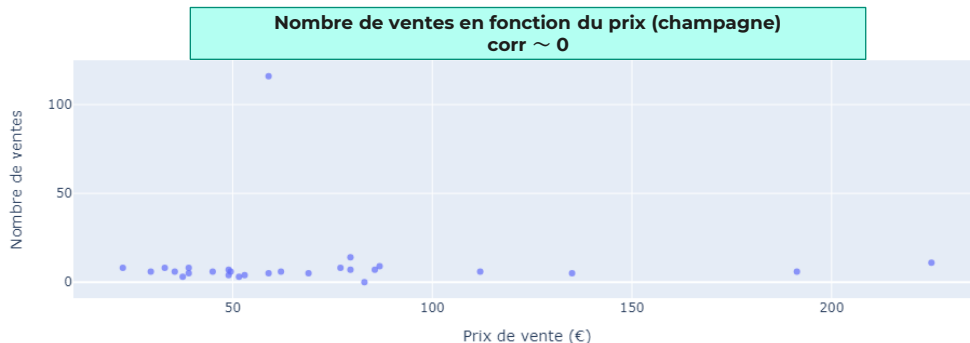
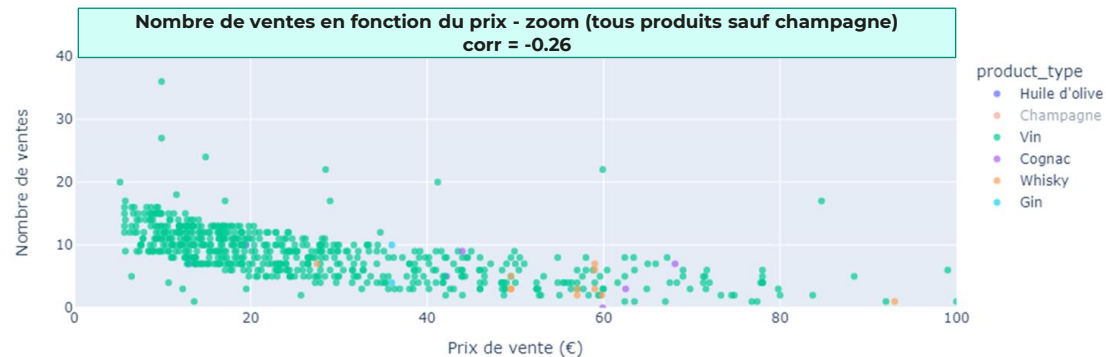
- On observe de fortes différences entre les heatmaps de corrélation par type de produit
- La forte corrélation négative observée entre marge et stocks est invalidée par la différenciation par type de produit
- **Certaines caractéristiques propres aux produits (saisonnalité, périssabilité, standing ...), peuvent justifier une stratégie marketing et logistique différenciée**



6. Analyse multivariée



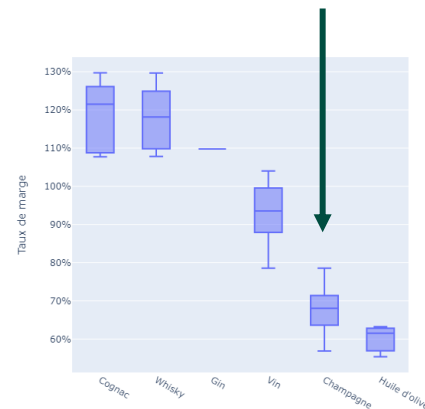
6.1 Corrélation prix et nombre de ventes : nuages de points



Validation des deux hypothèses :

- Courbe décroissante du nombre du vente en fonction du prix
- Nombre de ventes de champagne constante en fonction du prix

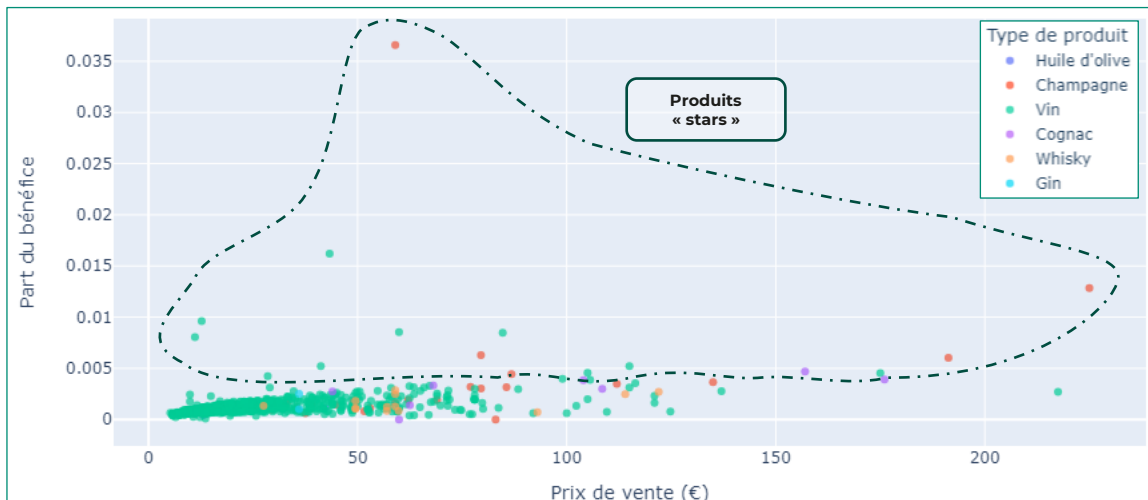
On rappelle que la marge est faible sur le champagne : **il pourrait être intéressant d'étudier comment une augmentation des marges impacterait les ventes sur le champagne**





6.2 Corrélation entre part du bénéfice et prix de vente : nuage de points

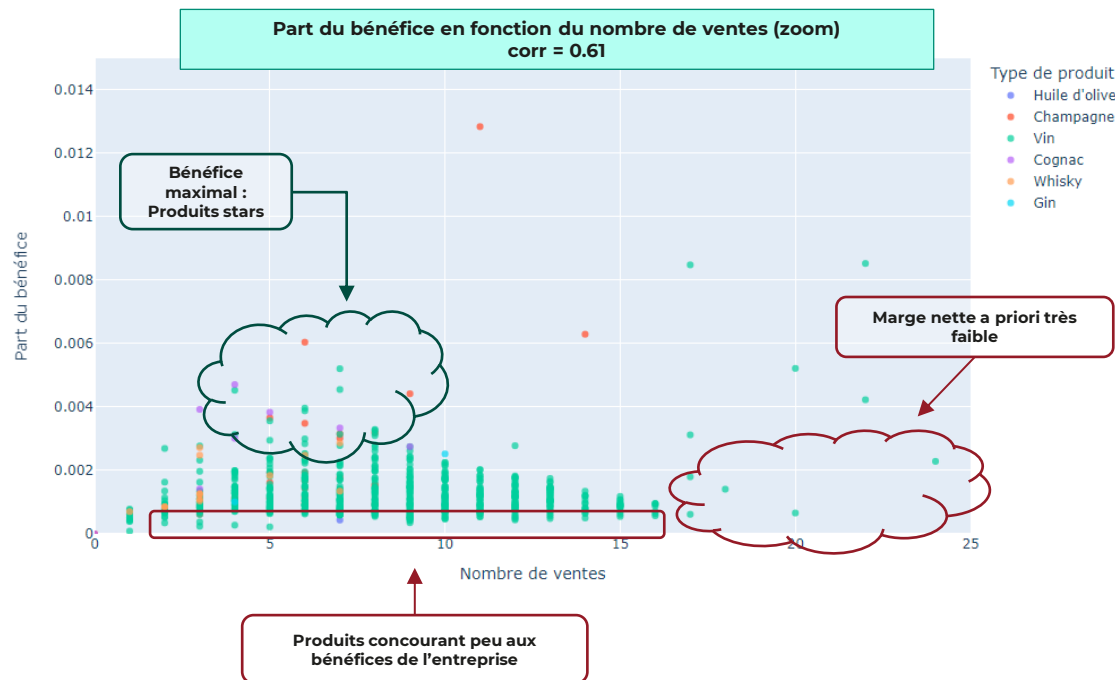
Part du bénéfice en fonction du prix de vente - corr = 0.37



- On valide une corrélation positive entre part du bénéfice et prix de vente
- Des produits bon marché comme des produits plus chers se démarquent comme produits « stars » au niveau du bénéfice des ventes : **une attention particulière doit leur être portée**



6.3 Corrélation entre part du bénéfice et nombre de ventes : nuages de points



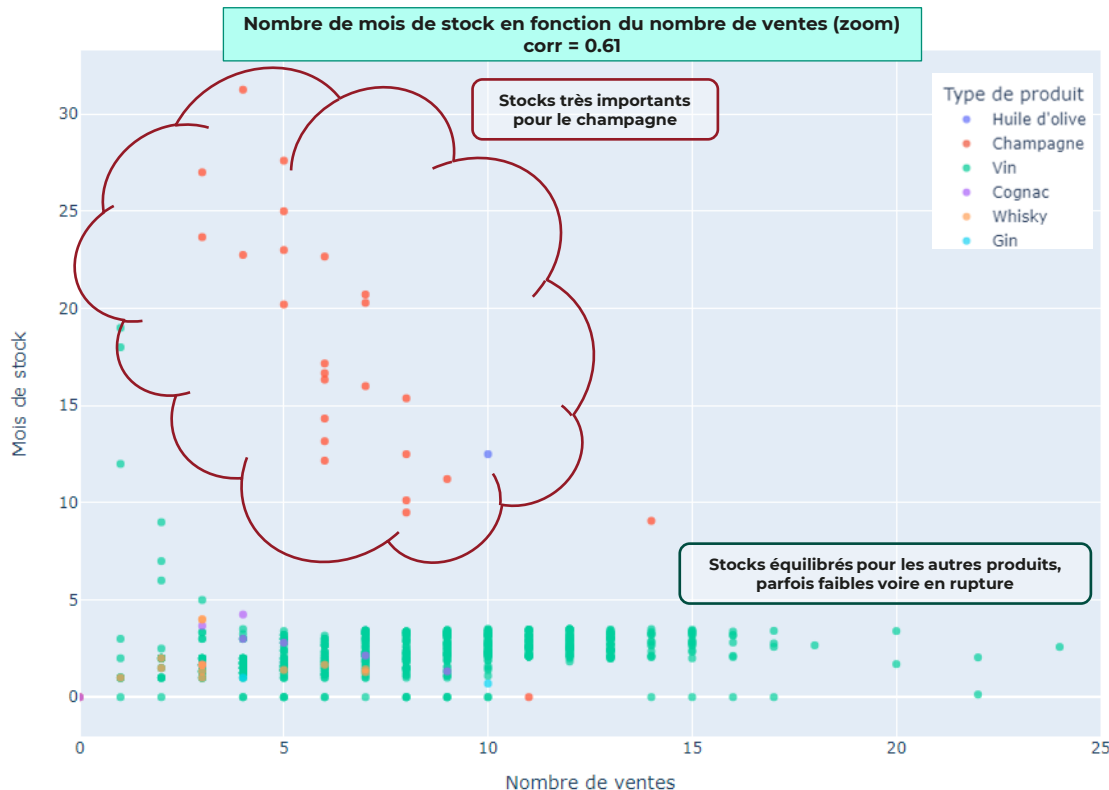
- La corrélation peu linéaire : on note que le **bénéfice maximal est obtenu pour les produits vendus autour de 6 articles par mois**

On propose donc les préconisations suivantes :

- **Augmenter la marge des produits à faible marge très vendus**, qui participent peu au bénéfice
- **Se concentrer sur les produits qui participent le plus au bénéfice** : seuil de part du bénéfice à définir afin d'analyser les produits situés en deçà?
- Limite de l'analyse : On travaille sur une **marge brute** ne prenant pas en compte ni les coûts fixes, ni les coûts linéaires de stockage et de main d'œuvre (plus pertes etc)
- Les produits les plus vendus sont ainsi favorisés dans cette analyse
- **Il serait pertinent de reprendre l'analyse avec la marge nette des produits**



6.4 Corrélation entre stock et nombre de ventes : nuage de points



- Tous les produits présentent environ 2 mois de stock, ce qui semble cohérent
- Les ruptures de stock seraient à analyser afin de les éviter au maximum
- Une règle pourrait être fixée pour uniformiser au maximum le nombre de mois de stock de chaque produit (vers 2 mois?)
- **Les stocks de champagne sont en revanche très importants et représentent un volume financier conséquent** : il conviendrait de vérifier si cet état de fait est justifié par la stratégie de l'entreprise



ERP

- Rechercher l'origine des prix négatifs et le cas échéant lever une erreur
- Rechercher l'origine des valeurs négatives du stock et le cas échéant lever une erreur
- Vérifier à chaque mouvement de stock que la valeur de 'stock_status' correspond à la valeur du stock enregistré
- **Remplacer la valeur 'onsale_web' (0/1) par l'identifiant du site 'id_web' -> la table de liaison devient inutile**

WEB

- **Ajouter un champ 'image' pour saisir le lien vers l'image du produit**
- Ajouter une clé étrangère 'product_id' correspondant à l'identifiant de l'ERP
- **Ne permettre la création que d'un produit par identifiant 'id_web' (sku)**
- Empêcher la saisie d'un produit sans identifiant conforme (nombre entier) avec un identifiant nul
- Chercher l'origine des nombre de ventes négatifs
- Vérifier si les 3 produits en double sont réellement les même, le cas échéant indiquer la différence
- Lever une alerte si un même nom de produit est déjà saisi

Analyse des données

- **Confirmer une stratégie de gestion des prix et des stocks différenciée entre les différents types de produits**
- Confirmer le choix des stocks (importants) et des marges (faibles) sur le champagne
- Rechercher l'origine des ruptures de stock
- **Consolider un nombre de mois de stock à viser** pour tous les produits pour éviter une immobilisation trop importante des capitaux et éviter les ruptures de stock
- Confirmer la vente de produits rapportant peu de bénéfice
- **Porter une attention aux produits qui se vendent bien** (autour de 6 par mois) par rapport aux produits chers peu vendus : **statistiquement, ils rapportent plus de bénéfice**
- Ajuster la marge des produits très vendus qui rapportent peu

Réaliser une analyse prenant en compte la marge nette des produits (marge brute – coûts fixes et coûts variable)



Points d'attention sur le nettoyage de données

- Enregistrement des éléments nettoyés : méthode d'identification et stratégie de nettoyage à consigner de manière claire et compréhensible
- Difficulté à choisir le traitement des valeurs aberrantes sans plus d'information : forme des échanges à avoir avec les équipes, synthèse à présenter au Codir

Points à approfondir

- Présentation concise des erreurs relevées lors du nettoyage
- Analyse bivariable : comment exploiter au mieux la grille de corrélation
- Méthodes statistiques d'analyse de données, de manière générale

Éléments positifs de la démarche

- Affichage des données sous forme de boîte à moustache très pratique pour visualiser clairement la forme des données
- Identification rapide des données aberrantes