Optimisez la gestion du stock d’une boutique en nettoyant ses données

wafa zargouni

[Data Analyst Et Business Intelligence]

[Date de la présentation]

Analyses Exploratoires des Données

- **Dataset: “erp.xlsx” :**

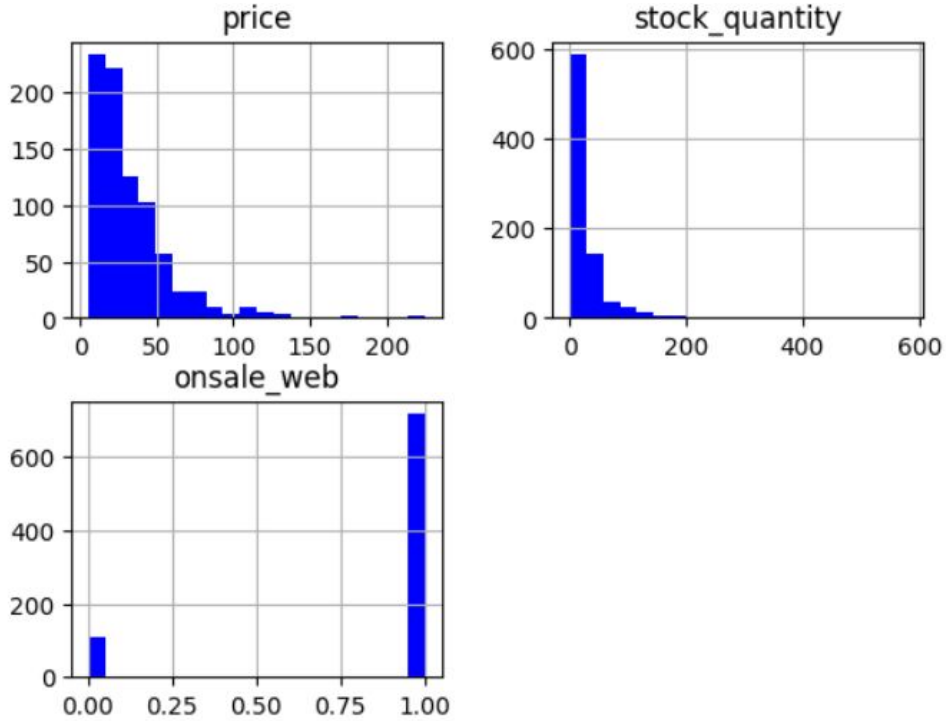
❖ *Caractéristiques :*

=> Dimensions (825, 4)

-> 825 observation(s) ou article(s)
-> 4 colonne(s)

=> Types des données :

-> float64,
-> INT,
-> Object (text)



Analyses Exploratoires des Données

- ***Traitement réalisés :***

❏ Vérification des doublons: toutes les lignes sont distinctes;
❏ Vérification des prix: tous les prix sont renseignés;
❏ Analyse de la variable “ONSALE_WEB”:
-> 87% articles vendus sur le web;
-> 13% pas vendus sur le web;
❏ Feature Selection:
-> suppression de la colonne “stock_status” (information redondante);
-> garder les colonnes: product_id, onsale_web, price, stock_quantity;

Analyses Exploratoires des Données

- **Dataset “web.xlsx” :**

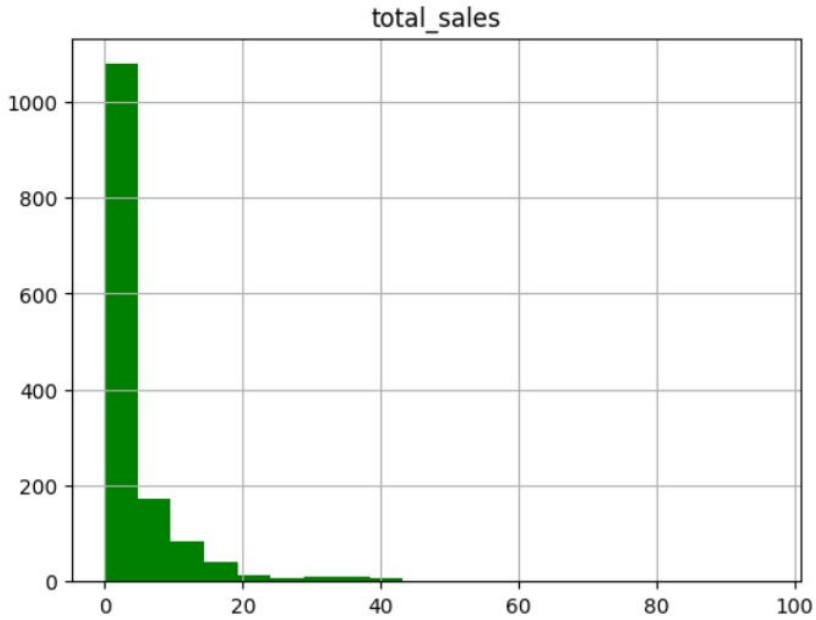
❖ *Caractéristiques :*

=> dimensions (1509, 24)

-> 1509 observation(s) ou article(s)
-> 24 colonne(s)

=> Type des données :

-> float64,
->INT,
->Object,
-> Datetime64[ns]



Analyses Exploratoires des Données

- ***Traitement réalisés :***

❏ vérification des doublons: 53% des lignes sont doublées;
❏ vérification des lignes sans clés primaire (“sku”): 85 non renseignées;.
❏ Handling outliers: 4 lignes avec des valeurs de “sku” qui ne respectent pas la règle de codification: ‘bon-cadeau-25-euros’ et ‘13127-1’

-> analyse: vérifier par “post_name”, supprimer le ‘-’ et vérifier;
-> action: supprimer ces 4 lignes.

❏ Feature Selection:
-> supprimer les colonnes vides: ‘tax_class', 'post_content', 'post_password', 'post_content_filtered’;
-> vérification des caractéristiques des lignes sans “sku”: pas de données
-> supprimer les lignes sans valeurs de “sku” renseignées

Analyses Exploratoires des Données

- **Dataset “liaison.xlsx” :**

❖ *Caractéristiques :*

=> Dimension du dataset: (825, 2)
-> Nombre d'observations: 825
-> Nombre de caractéristiques(colonnes) : 2
=> Type des données :
-> product_id: int64
-> id_web: object

❖ *Traitement réalisés :*

=> *vérification des doublons:*
-> pas de duplication de la colonne 'product_id'
-> 90 duplications de 'id_web'
=> vérification des articles sans correspondances c'est: 91
=> supprimer ces articles

Analyses Exploratoires des Données

- ***Dataset "caractéristiques vins.csv":***

❖ *Caractéristiques*

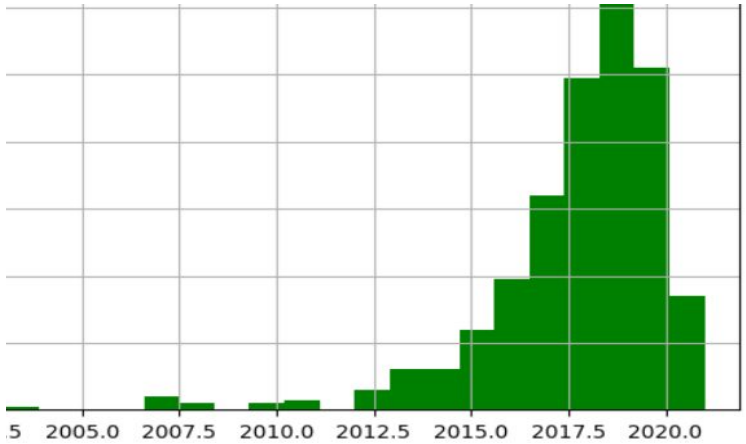
=> Dimension du dataset (611, 13)

-> Nombre d'observations 611

-> Nombre de caractéristiques 13

=> Types des données:

-> float64, object



❖ *Traitement réalisés*

=> vérifier les produits avec des informations manquantes

-> difficile de corriger les données manquantes

Fusion ou consolidations des données

- Jonction du fichier **“df_erp**” et **“df_liaison”** :

  => Choix des attributs: Primary Key
  =>Clé utilisé : product_id
  => Toutes les lignes ont des correspondances

- Jonction du fichier” **df_merge**” et **“df_web”** :

  =>Choix des attributs: Primary Key
  =>Clés utilisés : id_web, sku
  => puisque on a déjà supprimé les article sans “sku” dans “df_web”
  maintenant on n’a plus des lignes sans correspondances

- Jonction du fichier “**df_merge”** et “**df_caracteristiques”**:

  =>Choix des attributs: Primary Key
  =>Clé utilisé: post_name
  => toutes lignes ont des correspondances

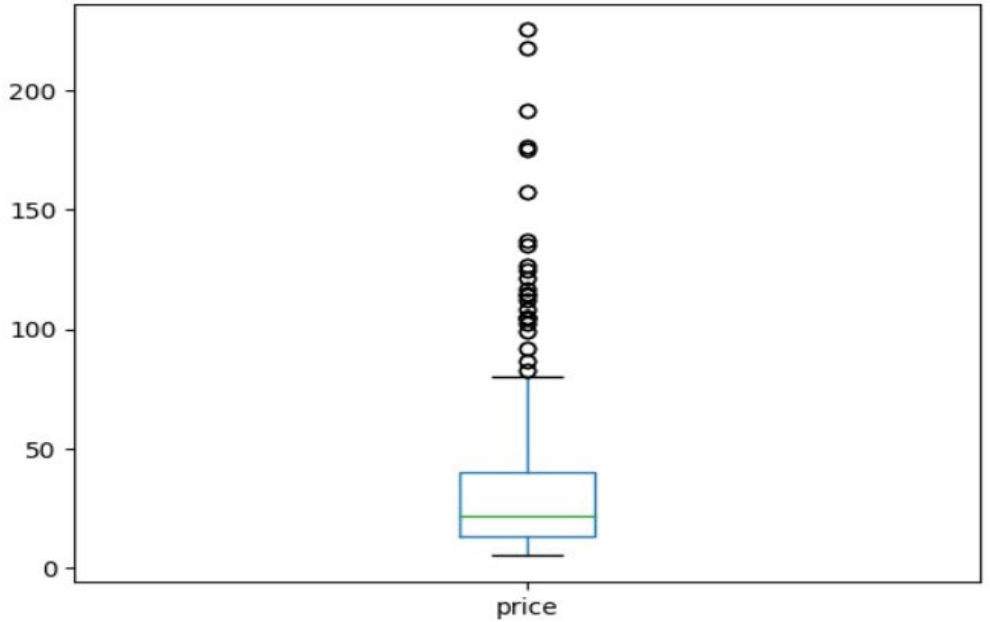
Analyses univariées du prix

- **Exploration par la visualisation de données**

-> Utilisation d'une boîte à moustache de la répartition des prix (boxplot)

=> Seuil ~ 80

=> Limite: Seuil pas exacte



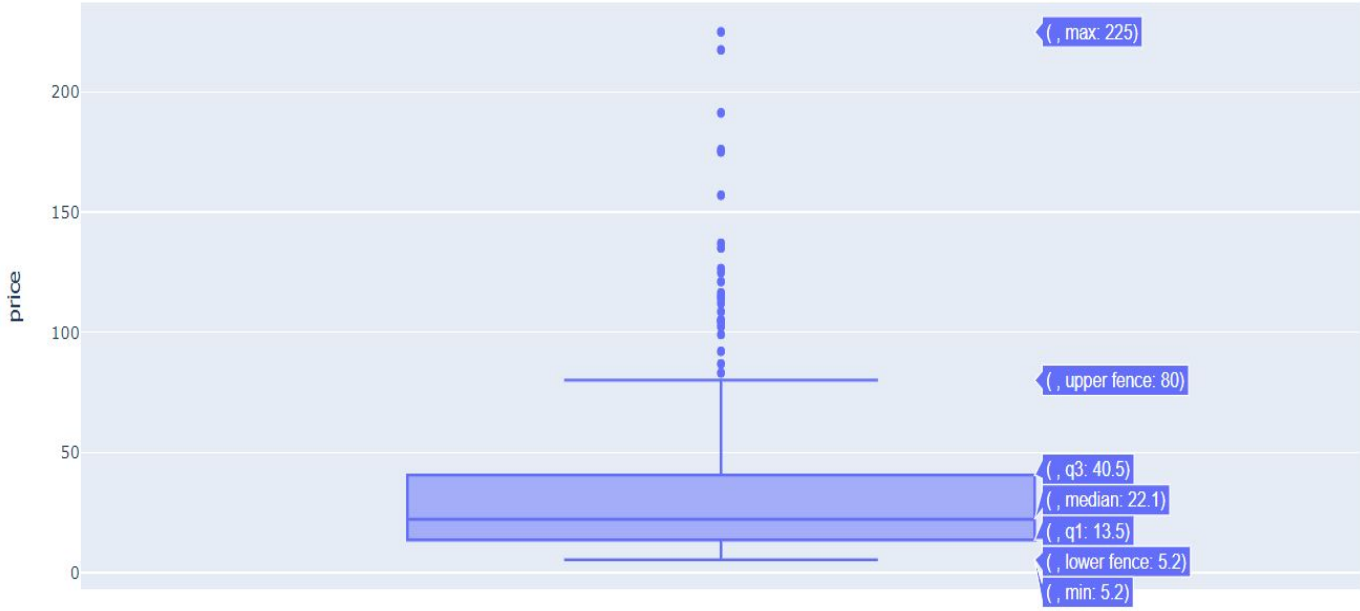
Analyses univariées du prix

- **Méthode avec plotly express(box)**

=> upper fence: 80

➢ Remarques:
-> Facile
-> valeur exacte

➢ Limites:
-> seuil fix



Analyses univariées du prix

- **Méthodes statistiques employés**

❖ Identification par le Z-index

➢ Z-index = (X – μ) / σ

X = total_prices

μ = prix_moyen

σ = standardDeviation(écart-type)

```
df_merge['z_score'][:5]
```

```
0   -0.256916
1   -0.256916
2    0.105010
3    0.105010
4   -0.378752
Name: z_score, dtype: float64
```

l'interval interquartile (describe)

```
count    1222.000000
mean       31.369558
std        27.917707
min         5.200000
25%        13.500000
50%        22.100000
75%        40.425000
max       225.000000
Name: price, dtype: float64
```

Analyses univariées du prix

- ❏ *Définition d'un seuil pour les articles "outliers" en prix*
  - *-> z-score min 2*
  - *-> Seuil prix: 92*
- ❏ Nombre d'articles outliers 48
- ❏ outliers justifiés
- ❏ vins de lux

- ➢ Remarques
  - -> valeurs exactes
  - -> flexible: selon z-score min

- ➢ Limites
  - ->choix du z-score min
  - ->calcul manuel difficile

post_name	price
champagne-egly-ouriet-grand-cru-millesime-2008	225.0
champagne-egly-ouriet-grand-cru-millesime-2008	225.0
david-duband-charmes-chambertin-grand-cru-2014	217.5
david-duband-charmes-chambertin-grand-cru-2014	217.5
coteaux-champenois-egly-ouriet-ambonnay-rouge-2016	191.3
coteaux-champenois-egly-ouriet-ambonnay-rouge-2016	191.3
cognac-frapin-vip-xo	176.0
cognac-frapin-vip-xo	176.0
camille-giroud-clos-de-vougeot-2016	175.0

Analyses univariées du CA

- **Méthodes statistiques employés**

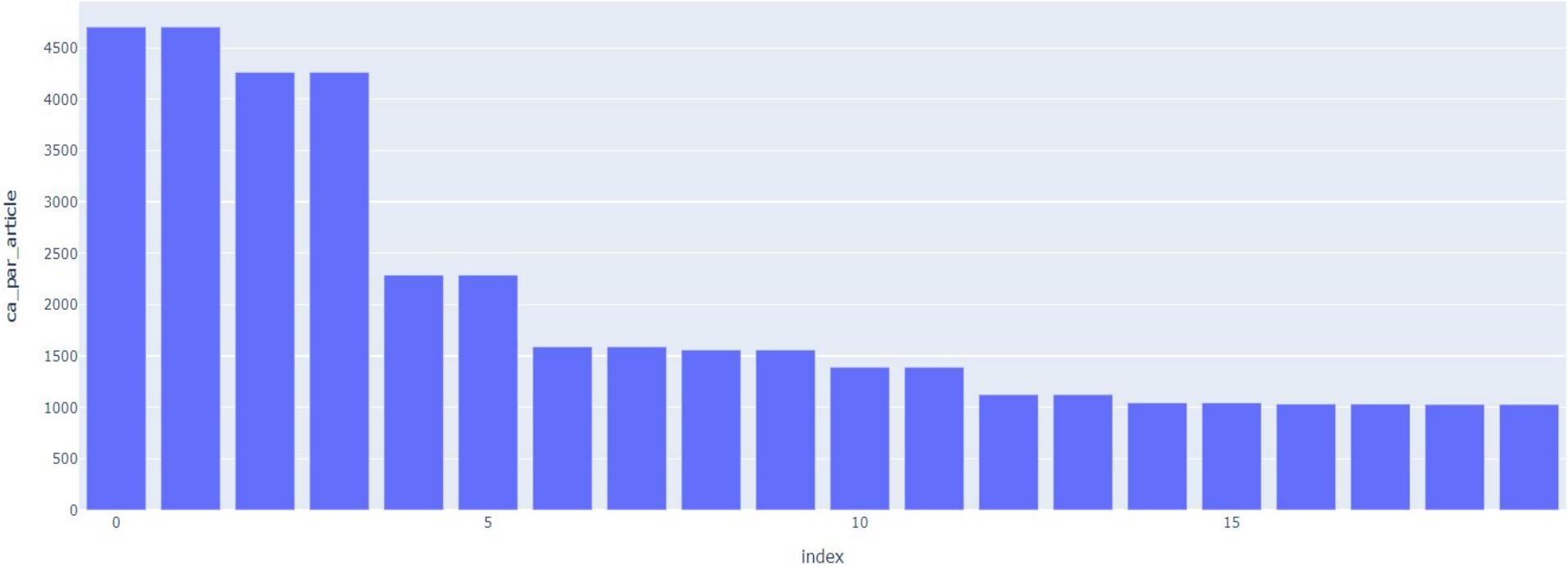
❖ *Analyse des ventes en CA*

Etapes:

=> calcul de CA par article
=> calcul de CA total (de site web): 127 274 €
=> Tri dans l'ordre décroissant du CA
=> calcul des sommes cumulatives de CA
=> calculer le nombre d'articles représentant 80% du CA: 227
=> ce groupe représente 18.5 % du catalogue entier du site web

Analyses univariées du CA

=> Graphique en barre des 20 premiers articles



Analyses univariées du CA

- **Méthodes statistiques employés**

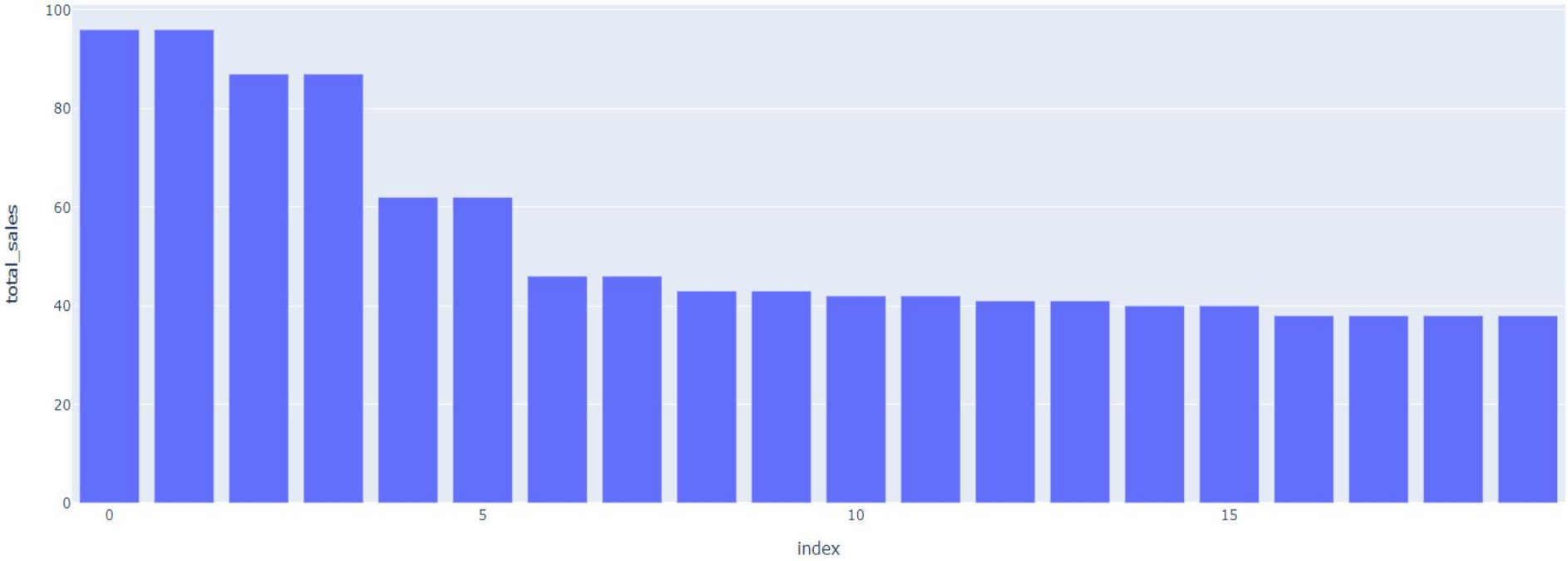
❖ Analyse des ventes en Quantités

Etapes:

=> tri dans l'ordre décroissant de quantités vendues
=> calculer la part en quantité
=> calculer les sommes cumulatives
=> calculer le nombre d'articles représentant 80% des ventes en quantité: 269
=> ce groupe représente 22% du catalogue entier du site web

Analyses univariées du CA

=> Graphique en barre des 20 premiers articles



Actions pour la suite

=>Exporter le dataset en fichier Excel

B	C	D	E	F	G	H	I	J	K	L	M	N
index	product_id	post_name	price	stock_quantity	z_score	ca_par_article	percent_ca	ca_percent_cum	ercent_total_sale	ent_total_sales_	Alliance mets_y	id_web
363	4334	champagne-gosset-grand-blar	49	0	0,63177324	4704	3,695939809	3,695939809	1,832760596	1,832760596	Apéritif, Coquille	7818
362	4334	champagne-gosset-grand-blar	49	0	0,63177324	4704	3,695939809	7,391879618	1,832760596	3,665521191	Apéritif, Coquille	7818
124	4144	champagne-gosset-grand-rose	49	11	0,63177324	4263	3,349445452	10,74132507	1,66093929	5,326460481	Apéritif, Dessert	1662
125	4144	champagne-gosset-grand-rose	49	11	0,63177324	4263	3,349445452	14,09077052	1,66093929	6,987399771	Apéritif, Dessert	1662
57	4068	gilles-robin-crozes-hermitage-	16,6	157	-0,52925568	1029,2	0,8086439735	31,47190174	1,183657885	8,171057656	Apéritif, Charcut	16416
56	4068	gilles-robin-crozes-hermitage-	16,6	157	-0,52925568	1029,2	0,8086439735	30,66325777	1,183657885	9,35471554	Apéritif, Charcut	16416
218	4200	moulin-de-gassac-igp-pays-dh	5,8	190	-0,91626532	266,8	0,2096251575	60,06247898	0,8781977854	10,23291333	Apéritif, Grillade	16295
219	4200	moulin-de-gassac-igp-pays-dh	5,8	190	-0,91626532	266,8	0,2096251575	60,27210414	0,8781977854	11,11111111	Apéritif, Grillade	16295
176	4172	maurel-pays-oc-chardonnay-2	5,7	167	-0,91984874	245,1	0,1925754352	62,06138214	0,8209240168	11,93203513	Apéritif, Poisson	16210
177	4172	maurel-pays-oc-chardonnay-2	5,7	167	-0,91984874	245,1	0,1925754352	62,25395758	0,8209240168	12,75295914	Apéritif, Poisson	16210
199	4187	le-pas-de-lescalette-languedo	13,3	90	-0,64750863	558,6	0,4388928523	48,00478964	0,8018327606	13,55479191	Agneau, Charcut	16189
198	4187	le-pas-de-lescalette-languedo	13,3	90	-0,64750863	558,6	0,4388928523	47,56589678	0,8018327606	14,35662467	Agneau, Charcut	16189
1121	6206	domaine-giudicelli-patrimonic	25,2	120	-0,22108133	1033,2	0,8117867795	29,04282702	0,7827415044	15,13936617	Fruits de mer, La	16580
1120	6206	domaine-giudicelli-patrimonic	25,2	120	-0,22108133	1033,2	0,8117867795	29,8546138	0,7827415044	15,92210767	Fruits de mer, La	16580
121	4141	gosset-champagne-grande-res	39	1	0,27343098	1560	1,225694324	21,4103656	0,7636502482	16,68575792	Apéritif, Fruits c	304
120	4141	gosset-champagne-grande-res	39	1	0,27343098	1560	1,225694324	22,63605993	0,7636502482	17,44940817	Apéritif, Fruits c	304
593	4729	emile-boeckel-cremant-brut-b	8,6	151	-0,81592949	326,8	0,2567672469	56,70124801	0,7254677358	18,17487591	Apéritif, Fromag	38
1088	6047	chateau-de-la-liquiere-faugere	10,9	46	-0,73351077	414,2	0,3254375572	53,77985273	0,7254677358	18,90034364	Apéritif, Charcut	16264
1089	6047	chateau-de-la-liquiere-faugere	10,9	46	-0,73351077	414,2	0,3254375572	53,45441517	0,7254677358	19,62581138	Apéritif, Charcut	16264

Point sur les compétences apprises

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

  => vérification des caractéristiques (dimensions, types des données, ..)
  => vérification des doublons,
  => utilisation de graphique avec pyplot express ou box pour déterminer le seuil

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

  => correction des données manquantes
  => calcul manuel de z-score et écart-type

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

  => correction des données manquantes