Analyse des ventes en ligne de nos vins et alcools

Thierry Monjo
Business Intelligence analyst

Juin 2024

Analyses Exploratoires des Données

- *Création du DataFrame « **erp** » par import du fichier Excel erp.xlsx*
- ***Caractéristiques** : 825 lignes sur 6 colonnes*
- ***Traitements réalisés** :*
  - ***Nettoyage des données** : mise en cohérence de la variable « stock_status » et « stock_quantity »*
  - ***Features engineering** : correction de deux valeurs négatives « stock_quantity » en valeurs positives, correction des valeurs négatives de la variable « price »*
- ***A noter** : **validation des données** par mise en cohérence des variables et corrections des valeurs négatives inattendues* avant étude

Analyses Exploratoires des Données

- *Création du DataFrame « **web** » par import du fichier Excel web.xlsx*
- ***Caractéristiques** : 1513 lignes sur 29 colonnes, dont nombre de champs portant sur des commentaires*
- ***Traitements réalisés** :*
  - ***Nettoyage des données** : les lignes d'index « sku » (Stock Keeping Unit) pour une même variable « total_sales » apparaissent deux fois -> conservation d'une ligne sur deux*
  - ***Features engineering** : suppression des lignes avec un « sku » non valide et simplification du DataFrame -> 3 variables « sku » , « total_sales » et « product_type » sur 713 lignes conservées*
- ***A noter** : repérage et suppression des doublons et simplification du DataFrame*

Analyses Exploratoires des Données

- *Création du DataFrame « **liaison** » par import du fichier Excel liaison.xlsx*
- ***Caractéristiques** : 825 lignes sur 2 colonnes, 91 valeurs de correspondance « product_id » et « id_web » sont manquantes et seraient à compléter*
- ***Traitements réalisés** :*
  ***Features engineering** : changement de nom de colonne de « id_web » en « sku » afin de préparer la fusion des DataFrames à venir*
- ***A noter** : identification du champ commun entre liaison et web : « id_web » de liaison est « sku » de web*

Fusion ou consolidations des données

- *Création du DataFrame « erp_liaison » par **fusion** de « erp » et « liaison »*
- *erp_liaison = pd.merge(erp, liaison, how='outer', on='product_id', indicator=True)*

```
RangeIndex: 825 entries, 0 to 824
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   product_id      825 non-null    int64
 1   onsale_web      825 non-null    int64
 2   price           825 non-null    float64
 3   stock_quantity  825 non-null    int64
 4   stock_status    825 non-null    object
 5   purchase_price  825 non-null    float64
 6   sku             825 non-null    object
dtypes: float64(2), int64(3), object(2)
```

- ***A noter** : les 825 lignes générées sont communes*

Fusion ou consolidations des données

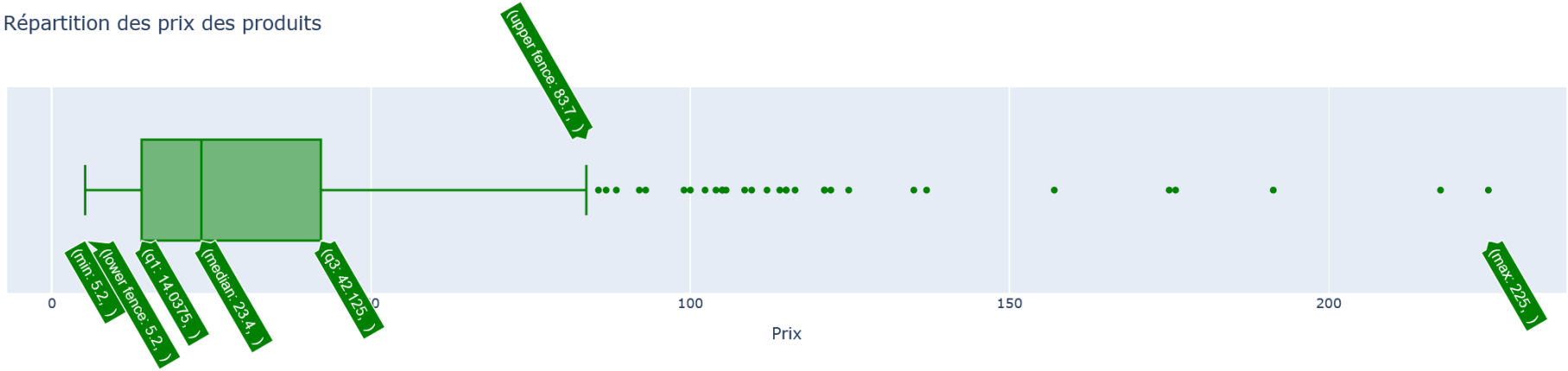
- *Création du DataFrame « produits_web » par **fusion** de « erp_liaison » et « web »*
- *produits_web = pd.merge(erp_liaison, web, how='outer', on='sku', indicator=True)*

```
Index: 713 entries, 0 to 824
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   product_id      713 non-null    int32
 1   onsale_web      713 non-null    int64
 2   price           713 non-null    float64
 3   stock_quantity  713 non-null    int64
 4   stock_status    713 non-null    object
 5   purchase_price  713 non-null    float64
 6   total_sales     713 non-null    float64
 7   product_type    713 non-null    object
```

- ***A noter** :*
  - *112 valeurs de erp_liaison n'ont pas donné lieu à des ventes -> suppression des lignes -> de 825 à 713 lignes*
  - *Correction d'une valeur « onsale_web » non cohérente -> de 0 à 1*

Analyses univariées du prix

- *Méthodes statistiques employées :*
  - *Exploration visuelle avec plotly express : boxplot*



  - *Limites de l'analyse : un certain nombre de prix « outliers », mais comment les évaluer ?*

Analyses univariées du prix

- *Méthodes statistiques employées :*
  - *Identification des valeurs aberrantes (outliers) par le **z-score** :*
    z-score = (x - moyenne) / écart-type

    Outliers ~ z-score > 3 (ou 2)

    13 valeurs retenues

	product_id	price
0	4352	225.0
1	5001	217.5
2	5892	191.3
3	4402	176.0
4	5767	175.0
5	4406	157.0
6	4904	137.0
7	6126	135.0
8	5612	124.8
9	5917	122.0
10	6213	121.0
11	6216	121.0
12	6202	116.4

  - *Limites de l'analyse : quel seuil retenir 2 ou 3 (34 valeurs retenues contre 13) ?*

Analyses univariées du prix

- *Méthodes statistiques employées :*
  - *Identification des outliers par **l'intervalle interquartile** : valeurs supérieures à la valeur du 3ème quartile plus 1.5 fois l'intervalle interquartile.*

    31 valeurs retenues

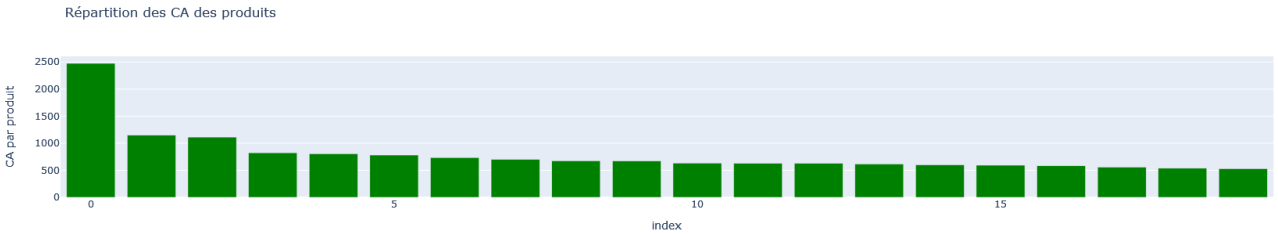
	product_id	price
0	4352	225.0
1	5001	217.5
2	5892	191.3
3	4402	176.0
4	5767	175.0
5	4406	157.0
6	4904	137.0
7	6126	135.0
8	5612	124.8
9	5917	122.0
10	6213	121.0
11	6216	121.0
12	6202	116.4
13	6212	115.0
14	6215	115.0
15	5918	114.0
16	5025	112.0
17	4582	109.6
18	4404	108.5
19	6201	105.6

Les 20 premières valeurs

  - *Limites de l'analyse : une méthode au spectre similaire au z-score > 2 une liste plus large de valeurs à contrôler*

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

- *Analyse des ventes en ligne en CA*
  - *Calcul du CA par produit (quantités vendues * prix de vente)*
    - *CA total = 143 505,10€*



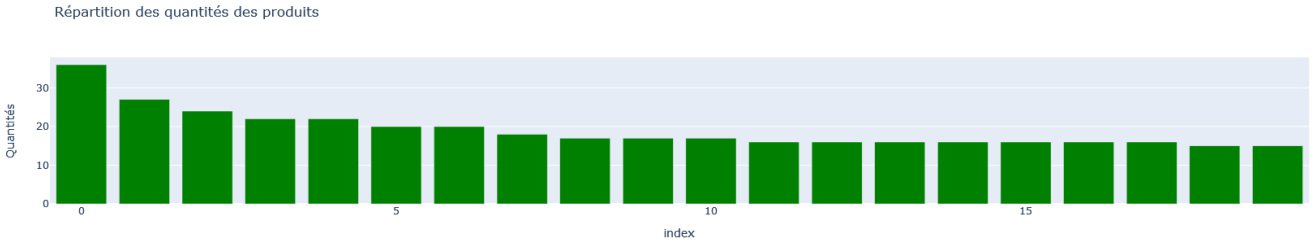
  - *Calcul des 20/80 du CA - principe de Pareto*
    nombre de produits qui représentent 80% du CA : 433
    (60,73% des produits)

	product_id	CA_prod
0	4352	2475.0
1	5892	1147.8
2	4353	1113.0
3	5826	824.0
4	6212	805.0
5	5026	781.2
6	5008	735.0
7	5767	700.0
8	6126	675.0
9	5025	672.0
10	6201	633.6
11	4406	628.0
12	4647	627.0
13	4358	616.0
14	4359	599.2
15	6214	594.0
16	6202	582.0
17	4350	556.5
18	4573	537.6
19	4402	528.0

Les 20 premières valeurs

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

- *Analyse des ventes en Quantités*
  - *Visualisation des quantités par produits*



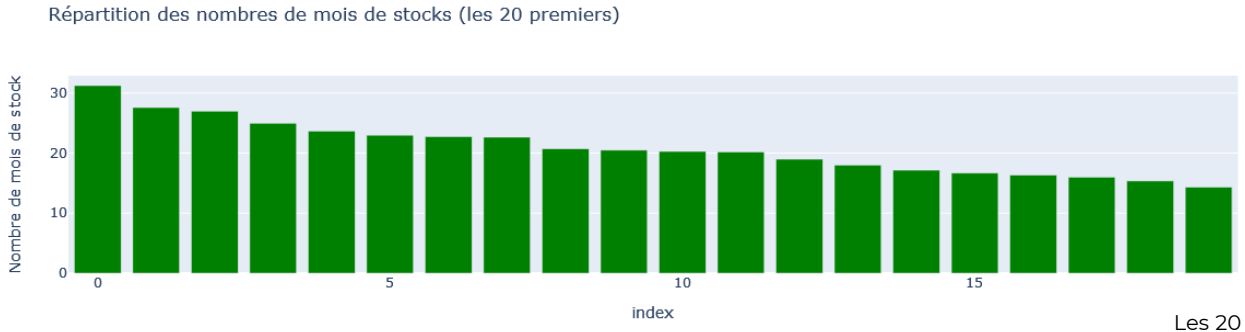
Les 20 premières valeurs

  - *Calcul des 20/80 du CA - principe de Pareto*
    nombre de produits qui représentent 80% des ventes : 432
    (60,59% des produits)

	product_id	total_sales
0	4867	36.0
1	4203	27.0
2	4275	24.0
3	4726	22.0
4	4647	22.0
5	5826	20.0
6	6129	20.0
7	4220	18.0
8	5778	17.0
9	6569	17.0
10	5803	17.0
11	4059	16.0
12	4188	16.0
13	4870	16.0
14	4105	16.0
15	5777	16.0
16	4863	16.0
17	5695	16.0
18	4261	15.0
19	4241	15.0

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

- *Analyse des stocks*
  - *Calcul des mois de stocks par produits*
    Principe :
    - *rotation des stocks = quantités mensuelles vendues/quantité de stocks*
    - *nombre de mois de stocks = 1/rotation de stocks*



Les 20 premières valeurs

	product_id	Nb_mois_stock
0	4142	31.25
1	6126	27.60
2	4356	27.00
3	4348	25.00
4	4148	23.67
5	4357	23.00
6	4144	22.75
7	5025	22.67
8	4350	20.71
9	4150	20.50
10	4334	20.29
11	4149	20.20
12	5612	19.00
13	4582	18.00
14	5024	17.17
15	4970	16.67
16	5892	16.33
17	4359	16.00
18	4141	15.37
19	4146	14.33

  - *Valorisation des produits en stock (approche comptable)*
    *Quantités en stock * valeur d'achat = 277 044,47€*

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

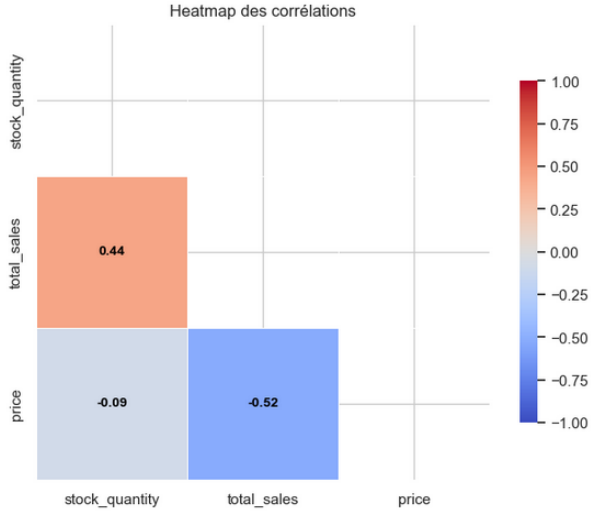
- *Analyse des taux de marge*
  - *Détermination du prix HT et du taux de marge*
    Principe :
    - *boissons alcoolisées -> taux TVA de 20%*
    - *Prix HT = prix de vente / 1,2*
    - *Taux de marge = (prix HT –prix d'achat)/prix HT*

	product_type	Taux_marge
0	Champagne	28.48
1	Cognac	45.07
2	Gin	42.80
3	Huile d'olive	25.00
4	Vin	38.01
5	Whisky	44.92

- ***A noter** : un Champagne avait une valeur de prix de vente erroné -> estimation à partir de la médiane des taux de marge des Champagnes*

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

- *Analyse des corrélations entre les variables stock_quantity, total_sales et price*
  - *Heatmap des corrélations*



Que peut-on en déduire ?

- Que quantités vendues et stocks sont **positivement corrélés** : plus les ventes sont importantes, plus les stocks sont importants
- Que quantités vendues et prix sont **négativement corrélés** : plus les prix sont élevés, moins les produits sont vendus
- Que les stocks et les prix des produits ne sont **pas corrélés** : les deux variables apparaissent indépendantes

Actions pour la suite

- *Une sauvegarde du DataFrame de travail **« produits_web »** au format Excel est réalisée pour partage des données (produits_web.xlsx)*

- *Pour rappel, le fichier Excel **liaison.xlsx est incomplet**, il manque 91 lignes de correspondance entre « product_id » et « id_web », il serait à compléter*

- *Un fichier Excel **outliers.xlsx** a été créé pour diffusion et contrôle auprès du service concerné, il reprend la liste élargie des 31 valeurs aberrantes relevée par la méthode de l'intervalle interquartile*

Point sur les compétences apprises

- *La validation des données : contrôles de cohérence, repérage et suppression des doublons, correction des valeurs anormales*
- *Le repérage des outliers ou valeurs aberrantes par deux méthodes différentes*
- *Les agrégations des données pour réaliser les regroupements clés (catégories) et leur visualisation*
- *La visualisation des corrélations entre les variables étudiées*