PROJET 6
Optimisez la gestion & nettoyez les données du stock d'une boutique

PAOLI J.
BUSINESS INTELLIGENCE ANALYST

Analyses Exploratoires des Données

Trois fichiers sources au format excel

- erp.xlsx
- web.xlsx
- liaison.xlsx

Un travail de nettoyage a été effectué sur les trois afin d’exploiter les données

Analyses Exploratoires des Données

- ERP

```
Le tableau comporte 6 colonne(s)
product_id          int64
onsale_web          int64
price             float64
stock_quantity      int64
stock_status       object
purchase_price    float64
dtype: object

nombre de valeurs par colonnes
product_id        825
onsale_web        825
price             825
stock_quantity    825
stock_status      825
purchase_price    825
dtype: int64
```

Analyses Exploratoires des Données

- ERP

- Correction des incohérences entre stock_quantity et stock_status : ajout d'une colonne stock_status_2 basée sur la quantité
- Vérification et correction des **prix aberrants**
- Détection et correction des **stocks négatifs**
- Suppression de la colonne stock_status_2 après vérification de la cohérence.

Analyses Exploratoires des Données

- WEB

```
Nombre de colonnes : 29

Types de données par colonne :
sku                          object
virtual                       int64
downloadable                  int64
rating_count                  int64
average_rating              float64
total_sales                 float64
tax_status                   object
tax_class                   float64
post_author                 float64
post_date                datetime64[ns]
post_date_gmt            datetime64[ns]
post_content                float64
product_type                 object
post_title                   object
post_excerpt                 object
post_status                  object
comment_status               object
ping_status                  object
post_password               float64
post_name                    object
post_modified            datetime64[ns]
post_modified_gmt        datetime64[ns]
post_content_filtered       float64
post_parent                 float64
guid                         object
menu_order                  float64
post_type                    object
post_mime_type               object
comment_count               float64
dtype: object
```

```
Nombre de valeurs présentes par colonne :
sku                     1428
virtual                 1513
downloadable            1513
rating_count            1513
average_rating          1430
total_sales             1430
tax_status               716
tax_class                  0
post_author             1430
post_date               1430
post_date_gmt           1430
post_content               0
product_type            1429
post_title              1430
post_excerpt             716
post_status             1430
comment_status          1430
ping_status             1430
post_password              0
post_name               1430
post_modified           1430
post_modified_gmt       1430
post_content_filtered      0
post_parent             1430
guid                    1430
menu_order              1430
post_type               1430
post_mime_type           714
comment_count           1430
dtype: int64
```

Analyses Exploratoires des Données

WEB

- Identification et suppression de colonne complément vide
- Identification et suppression de données incohérentes dans la colonne SKU
- Même opération concernant les colonnes sans SKU (les NaN)
- Vérification et suppression de lignes quasi identiques (Faux-doublons)

Analyses Exploratoires des Données

- Liaison

```
Nombre de colonnes dans df_liaison : 2

Types de données par colonne :
id_web         object
product_id      int64
dtype: object

Nombre de valeurs non nulles par colonne :
id_web         734
product_id     825
dtype: int64
```

Analyses Exploratoires des Données

Liaison

- Identification des valeurs uniques
- Et des valeurs manquantes

Analyses Exploratoires des Données

- Difficultés

La plus importante, hors fautes de frappes ou de remplissage; ce sont les colonnes quasi identiques de WEB qui ont failli fausser l’analyse

Fusion ou consolidations des données

🔗 **Le fichier liaison.xlsx est notre fichier de liaison**

Traitement réalisé :

- product_id et id_web sont nos **clés de liaison avec les bases erp et web respectivement**

- **Peu de traitements ont été efféctués mais principalement : suppression des lignes manquant d'information essentielles**

Analyses univariées du prix

🔧 **Méthodes statistiques employées**

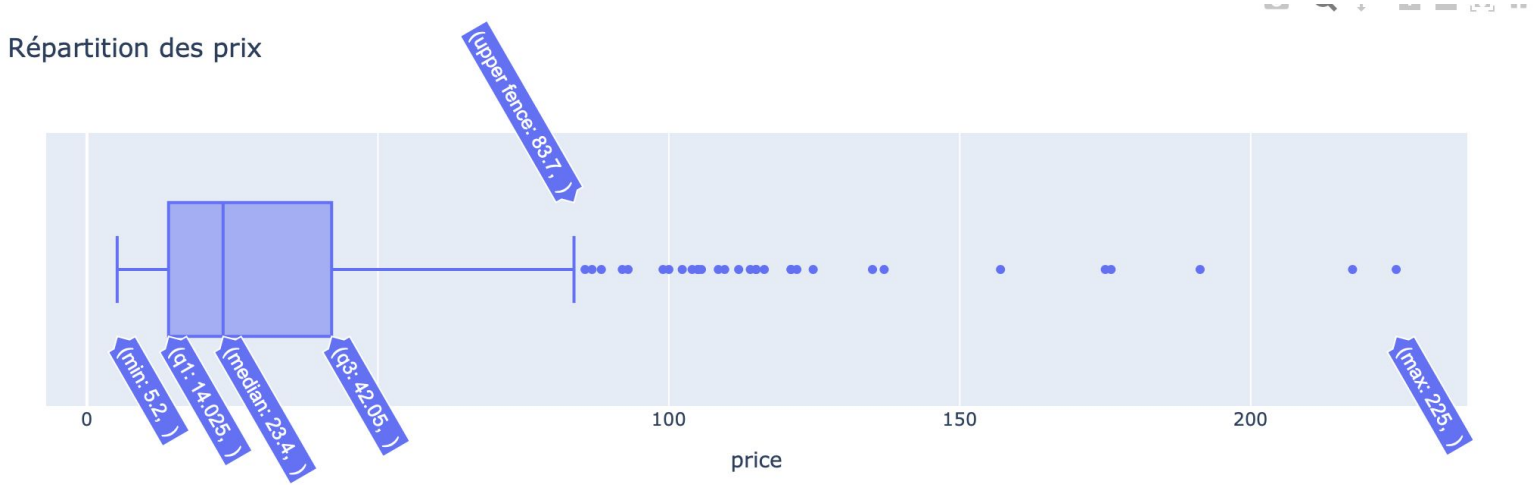
- **Boîte à moustaches (boxplot)** pour détecter visuellement les outliers

- **Statistiques descriptives** :
  - Moyenne, minimum, maximum, médiane (df['price'].describe())
  - Détection des valeurs aberrantes

- **Méthode IQR (Interquartile Range)** :
  - Calcul de Q1, Q3 et IQR
  - Détection des valeurs anormalement basses ou hautes avec les bornes et un histogramme :
    - borne_basse = Q1 - 1.5 × IQR
    - borne_haute = Q3 + 1.5 × IQR

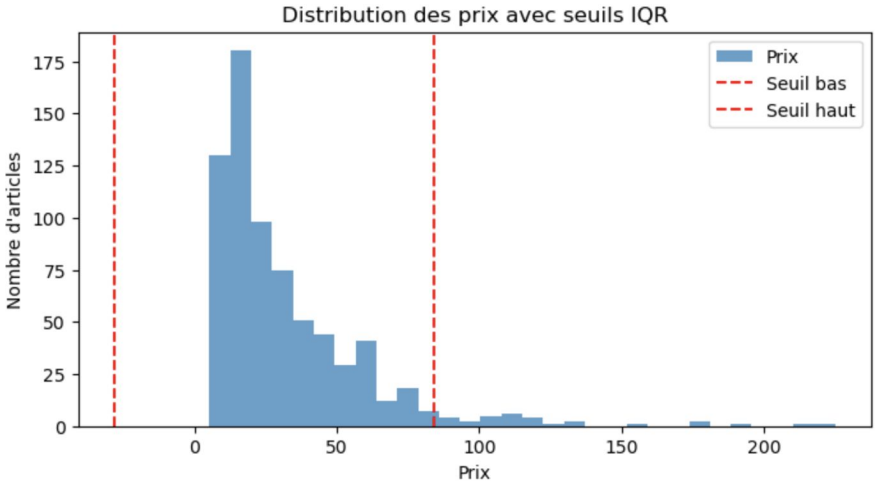
Analyses univariées du prix

📊 Graphiques



Analyses univariées du prix

📊 Graphiques = Méthode IQR



les valeurs abérantes sont apparament des bouteilles chères, ce qui est cohérent dans un magasin de spiritueux, et on remarque qu'aucune valeur n'est inférieur à zéro

Analyses univariées du prix

⚠️ **Limites de l’analyse**

- **Pas de distinction entre TTC et HT**
- La méthode IQR ne tient pas compte du contexte métier
- Aucun ajustement par catégorie de produit : les prix ne sont pas analysés **relativement à leur gamme**

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

🔧 Méthodes statistiques employées

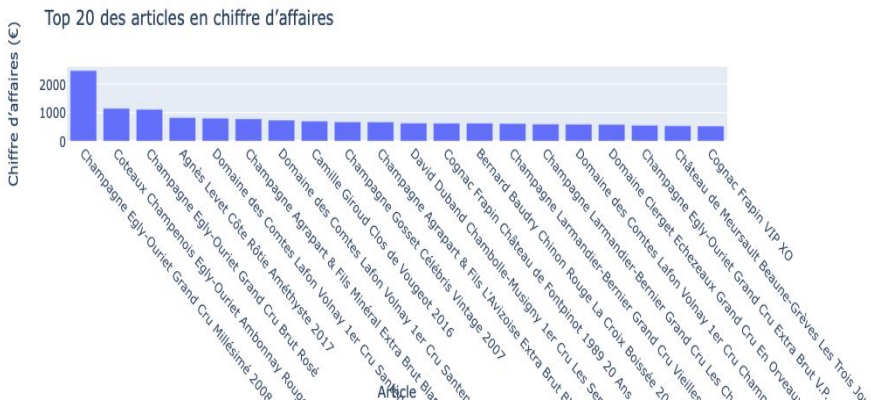
- **Calculs de base** :
  CA unitaire (price × total_sales), total CA, taux de marge, rotation de stock (stock_quantity / total_sales); Z-score
- **Analyse 20/80 (Pareto) :**
  Calcul de la part cumulée du CA ou des ventes pour identifier les articles clés
- **Statistiques descriptives** sur le taux de marge (min, max, distribution)

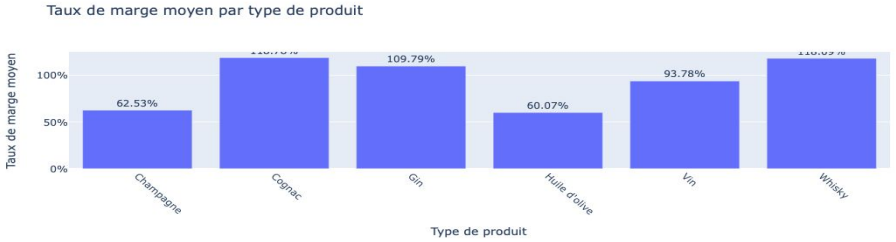
```
Résumé statistique de la variable 'price' :
count    712.000000
mean      32.312430
std       27.620894
min        5.200000
25%       14.037500
50%       23.400000
75%       42.025000
max      225.000000
Name: price, dtype: float64
```

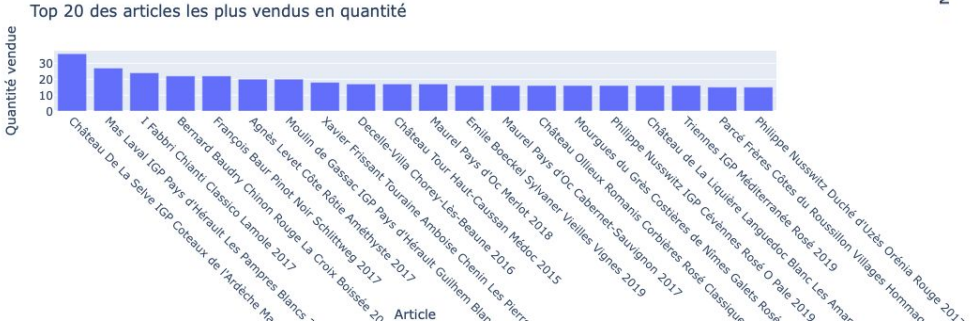
```
Chiffre d'affaires total du site web : 143285.9 €
```

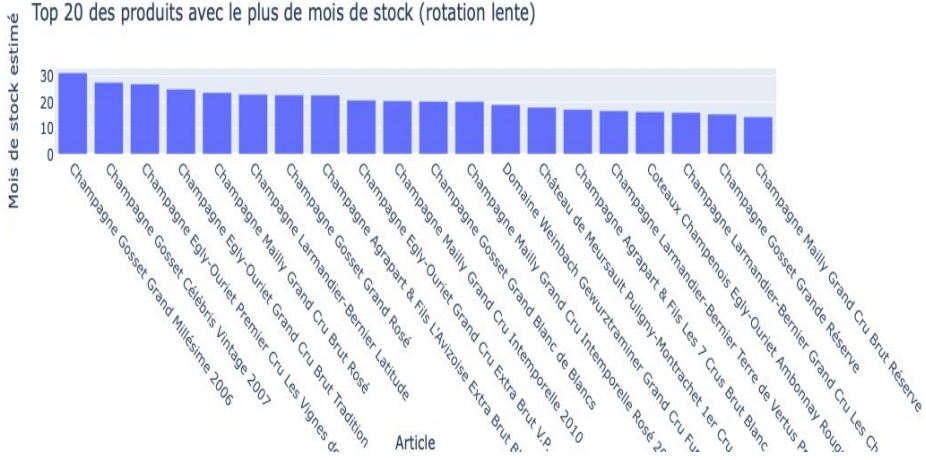
```
Nombre d'articles outliers (IQR) : 31
Proportion des outliers (IQR) : 4.35 %
```

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

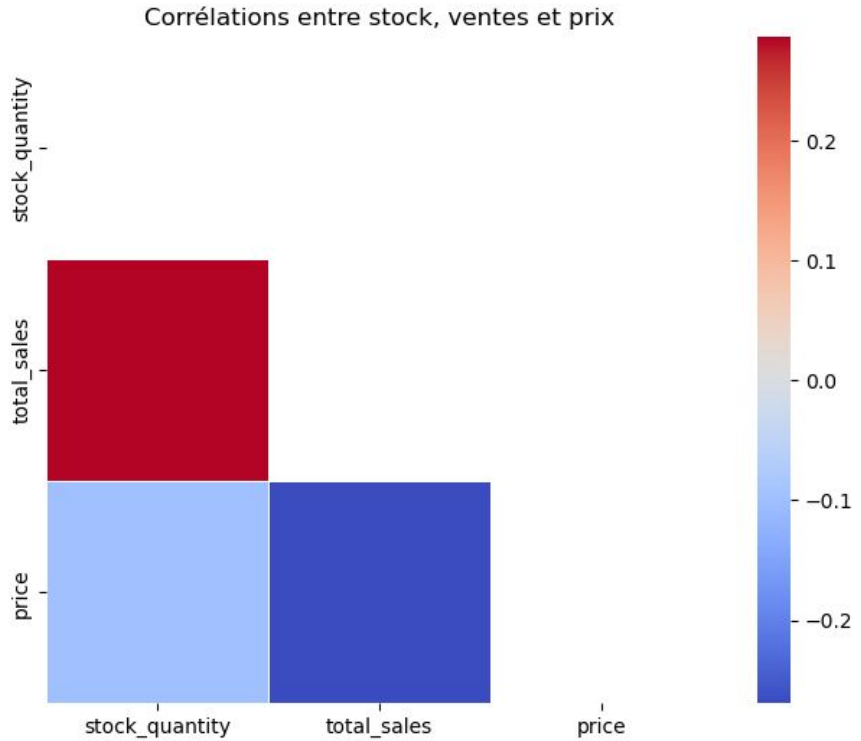








Analyses complémentaires
CA, quantités, stocks, taux de marge et correlations



Analyse de corrélation (matrice + heatmap Seaborn) : Entre stock_quantity, price, et total_sales

<u>TROIS TENDANCES</u>

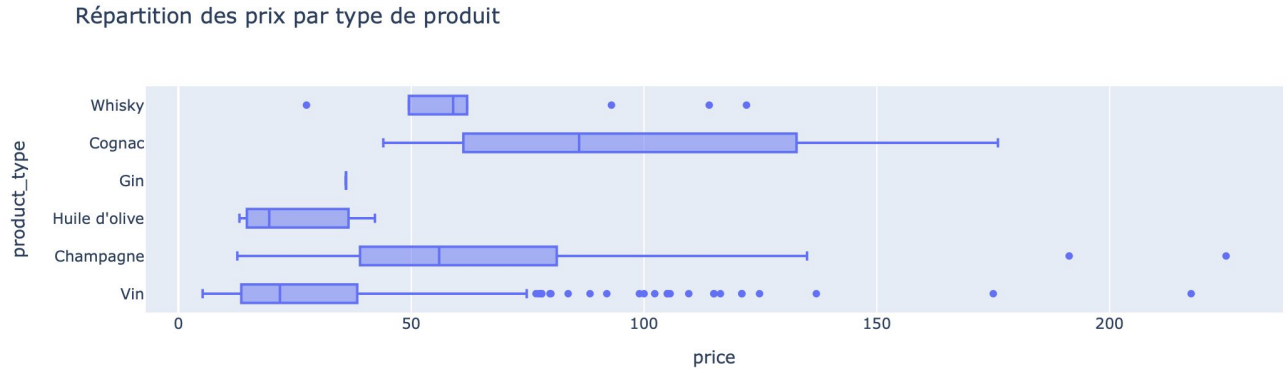
- Une corrélation positive entre les quantités en stock et les ventes
- Une corrélation négative entre le prix et les ventes
- Une légère corrélation négative entre le prix et le stock disponible

⚠️ Limites de l’analyse

- L’analyse de marge se fait sans savoir si les taxes sont appliquées ou pas
- Le stock est traité de manière linéaire
- Les données ne permettent pas d’analyser les promotions, remises ou retours
- Les nombreuses actions de nettoyage

Actions pour la suite

- Remplir les fichiers avec plus de soins
- Sélection des colonnes
- Éviter les Faux-doublons de lignes involontaires
- Diviser l'analyse pour plus de précision



Point sur les compétences apprises

- Le reflexe de vérification et de nettoyage des données
- La création de graphique notamment les boites à moustaches
- Les utilisations statistiques de Python

MERCI
POUR VOTRE ATTENTION