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

Chehrazad El Botout

28/10/2023

Analyses Exploratoires des Données

- Importation des librairies Pandas et Plotly Express
- Importation des fichiers web, erp, liaison et caractéristiques vins

Dataset : ERP

- Caractéristiques : 6 colonnes, 825 lignes
- Traitement réalisés :
- Nettoyage des données, analyse des variables, agrégation des données (max, min), vérification des doublons, des valeurs distinctes
- Features engineering : Création colonne stock status 2, corriger la ligne incohérente, suppression colonne inutile (stock status)

Analyses Exploratoires des Données

Dataset : WEB

- Caractéristiques : 28 colonnes, 1513 lignes
- Traitements réalisés :
- Nettoyages des données : Suppression des colonnes avec des valeurs nulles, identification des valeurs qui ne respectent pas la codification, verification des doublons, verification des informations manquantes
- Features engineering : Création Dataframe

Analyses Exploratoires des Données

Dataset : LIAISON

- Caractéristiques :

2 colonnes et 825 lignes, 825 valeurs dans ‘product_id’ et 734 valeurs dans ‘id_web’

- Traitement réalisés :

Nettoyages des données : Vérifier si les valeurs sont uniques, vérification des articles sans corespondances

Analyses Exploratoires des Données

Dataset : CARACTERISTIQUES VINS

- Caractéristiques : 13 colonnes et 611 lignes

- Traitement réalisés :

Nettoyages des données : Vérification des valeurs présentes dans chaque colonne, vérification des informations manquantes

Fusion ou consolidations des données

- Choix des attributs : ‘ How’ , pour déterminer le type de jointure
- Clés utilisés : left on, right on pour indiquer le df à associer
- Fonction utilisée : Merge, création du dataframe _merge pour effectuer les jointures entre les fichiers
- Jointure externe : Outer , inclut toutes les lignes des dataframes
- Valeur boléenne : True, pour inclure le résultat dans l’enregistrement
- Vigilances particulières au cours du traitement
- Difficultés ou pièges rencontrés

Analyses univariées du prix

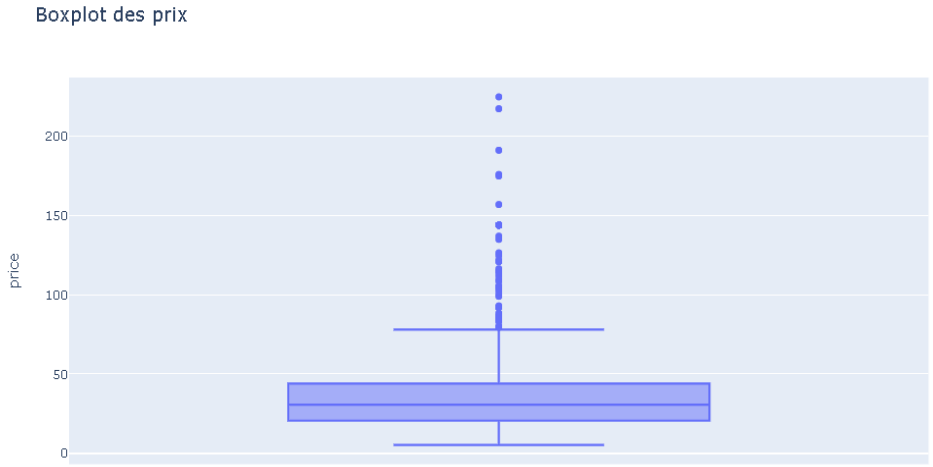
Méthodes statistiques employées

- Création d'une boite à moustache avec Panda et avec Plotly Express (Boxplot des prix)
- Calcul de la moyenne avec mean, de l'écart-type et du Z-score
- Intervalles interquartiles pour quantifier la dispersion des valeurs sur l'ensemble des données
  -> Q1 (25%), Q2(50%), Q3(75%), Max

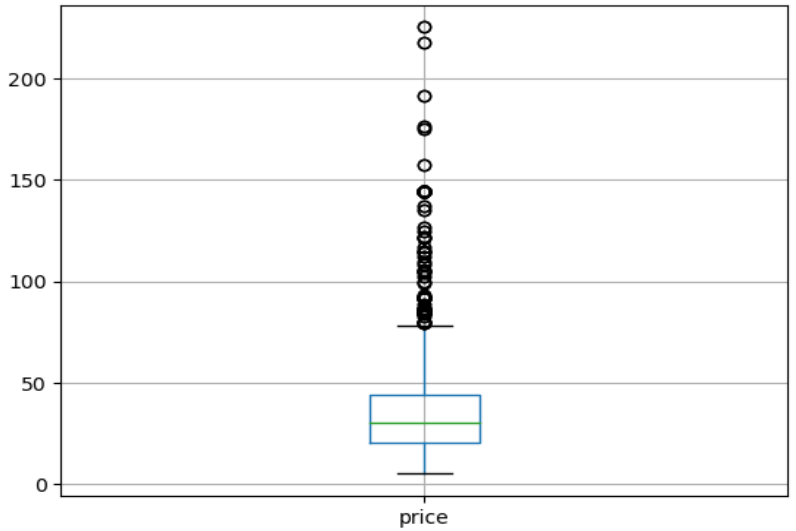
	price
count	825.000000
mean	32.415636
std	26.795849
min	5.200000
25%	14.600000
50%	24.400000
75%	42.000000
max	225.000000

Analyses univariées du prix

Boxplot avec Pandas



Boxplot avec Plotly Express



Analyses univariées du prix

Explication des graphiques de répartition des prix

- La boîte centrale représente la médiane, qui est la valeur médiane des données.
- La longueur de la boîte représente la dispersion interquartile.
- Les moustaches indiquent la plage de données en dehors de laquelle les valeurs sont considérées comme des valeurs aberrantes.
- Les points en dehors des moustaches sont des valeurs aberrantes.
- Le graphique est utile pour identifier la centralité, la dispersion, la symétrie et les valeurs aberrantes des données.

Analyses univariées du prix

Méthodes statistiques employées

- *La moyenne des prix est de 32.56*
- *L'écart-type des prix est de 27.86*
- *Le Z-score est de 1.16*
- *Le seuil prix pour le Z-score de 3 est : 116.4 €*
- *Les outliers se situent à 83.1 €*
- *Le nombre d'article au dessus du seuil des outliers est de : 37*
- *Proportion d'articles au dessus du seuil des outliers : 4%*

Analyses univariées du prix

Méthode utilisée pour comprendre la répartition des prix :

- *Seuil des Outliers (Valeurs aberrantes) : Le seuil outliers est à 83.1€*
- *Nombre d'articles au dessus des outliers : 37*
- *La moyenne des prix est de 32.54€*
- *L'ecart-type de prix est de 27.88 €*
- *Le Z-score est de 1.16*

Analyses univariées du CA

Méthodes statistiques employés

- *Chiffre d'Affaire du site web :* ***70 568,6 €***
- *Utilisation de sort_values pour trier par ordre décroissant le CA et les quantités vendues dans le dataset df_merge, puis utilisation de reset_index pour le réinitialiser.*
- *Utilisation de la fonction sum pour calculer le CA total.*
- *Division du CA de chaque ligne par le CA total pour obtenir la part du CA de chaque ligne*

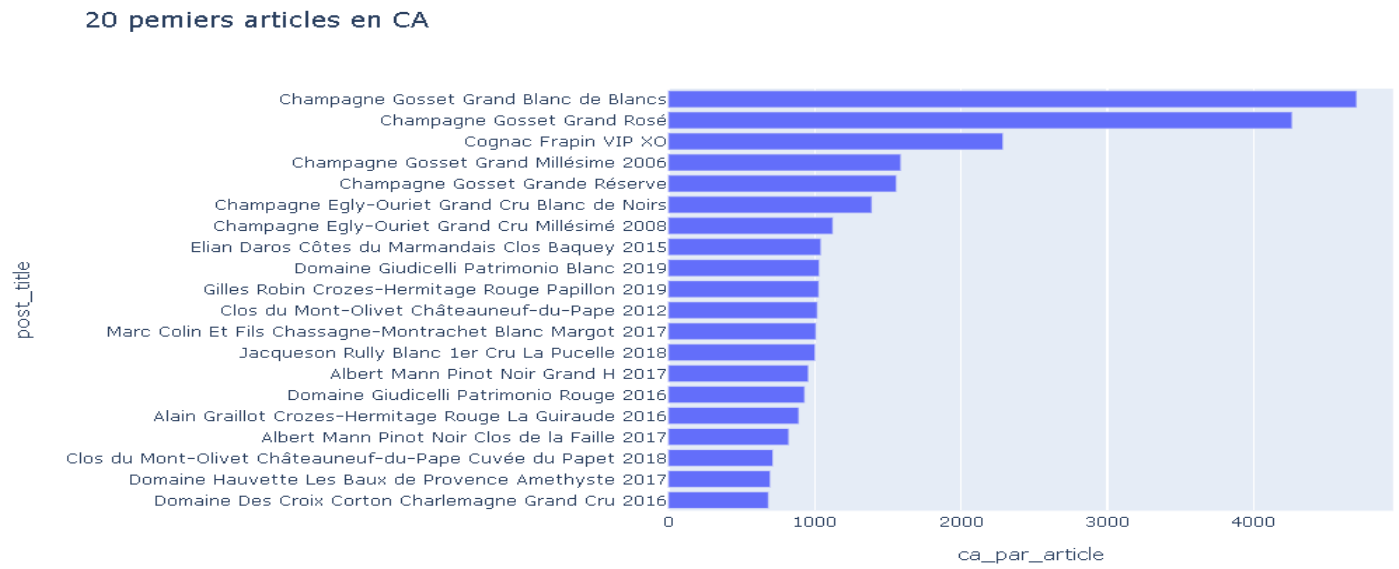
Analyses univariées du CA

Méthodes statistiques employés

- *Calcul du nombre d'articles représentant 80% du CA avec la fonction 'shape', ainsi que la proportion que représentent ces articles sur le catalogue du site web :* ***Il y a donc 130 articles qui représentent 18.18% du catalogue***
- *Utilisation de la même méthode pour calculer le nombre d'articles représentant 80% des articles en quantité, ainsi que la proportion qu'ils représentent :* ***Il y a donc 130 articles qui correspondent à 18.18% du catalogue.***

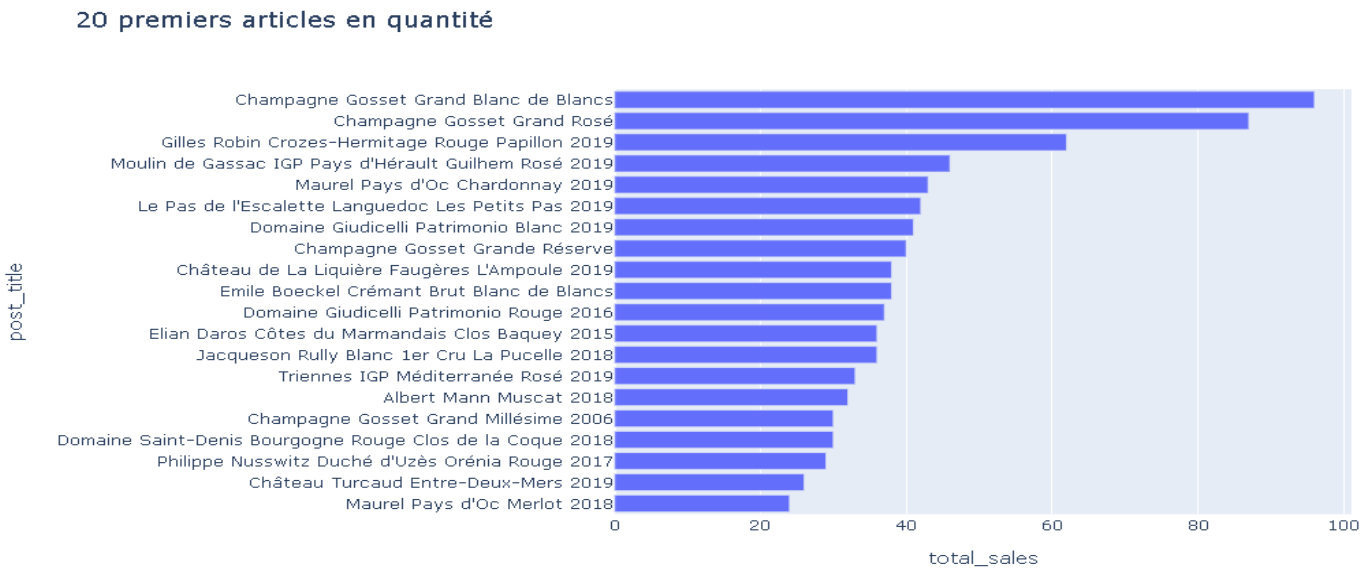
Analyses univariées du CA

Graphique en barre des 20 premiers articles classés par CA



Analyses univariées des quantités vendues

Graphique en barre des 20 premiers articles en quantité



Actions pour la suite

- Maintenir une bonne gestion des ressources, avec un site web mis à jour et ainsi avoir une bonne gestion du stock de l'entreprise
- Donner accès au back office à l'analyse des ventes pour accroitre les progrès de l'entreprise
- Analyser les données de ventes pour comprendre quelles catégories de vins se vendent le mieux et à quel moment de l'année

Point sur les compétences apprises

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

Le tri, les créations et les suppressions de colonnes.

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

La vérification de la cohérence des résultats et la fusion des fichiers

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

La correction des incohérences