bottleneck
fine wine spirit

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

Freger Morgan
Business Intelligence Analyse
27 Juin 2023

Analyses exploratoires des données

erp.xlsx

Caractéristiques :

- Nbr observations : 825 lignes
- Nbr variables : 5 colonnes
- Clés non renseignées : 0
- Doublons détectés : 0
- Clés uniques : 825

Traitements réalisés :

- Analyse variable 'stock_status' : erreur identifiée et corrigée (ligne 443)
- Variables 'stock_status' & 'stock_status_2' supprimées car redondance avec 'stock_quantity'

Remarques :

À première vue, la variable 'stock_status' paraissait correctement renseignée mais une erreur s'y était glisée

web.xlsx

Caractéristiques :

Nbr observations : 1513 lignes
Nbr variables : 28 colonnes
Clés non renseignées : 85
Doublons détectés : 798
Clés uniques : 714

Traitements réalisés :

- Suppression de 20 colonnes dont les informations n'étaient pas nécessaires à l'analyse.
- Identification et suppression des doublons de type 'attachment'
- Identification et suppression des lignes sans codes articles ('sku')

Remarques :

-Difficulté à trier les 28 colonnes pour conserver uniquement les variables pertinentes
- Difficulté dans le choix d'écarter ou non les 2 observations 'sans sku'

liaison.xlsx

Caractéristiques :

Nbr observations : 825 lignes
Nbr variables : 2 colonnes
Clés non renseignées : 91
Doublons détectés : 90
Clés uniques : 734

Traitements réalisés :
Pas de nettoyage réalisé à ce stade

Remarques :

- 91 observations n'ont pas de correspondances entre 'product_id' et 'id_web'

caractéristiques_vins.csv

Caractéristiques :

Nbr observations : 611 lignes
Nbr variables : 13 colonnes
Clés non renseignées : 0
Doublons détectés : 0
Clés uniques : 611

Traitements réalisés :
Pas de nettoyage réalisé à ce stade

Remarques :

- Erreur de lecture du fichier .csv (utilisation 'chardet.detect' pour déterminer encodage)

- 118 observations ont des données manquantes

Fusions ou consolidations des données

Jonction df_erp & df_liaison (= df_merge)

Choix des attributs :

.merge(on = 'product_id', how = "outer", indicator=True)

Clés utilisées :

'product_id' dans les 2 dataframes

Vigilances particulières au cours du traitements :

S'assurer de la correspondance des observations lors de la jonction

Difficultés ou pièges rencontrés :

Aucunes difficultés rencontrées, toutes les lignes matchent entre les 2 dataframes

Jonction df_merge & df_web (= df_merge)

Choix des attributs :

.merge(df_web, left_on = 'id_web', right_on = 'sku', how = 'left', indicator = True)

Clés utilisées :

'id_web' dans le dataframe 'df_merge'
'sku' dans le dataframe 'df_web'

Vigilances particulières au cours du traitements :

S'assurer de la correspondance des observations lors de la jonction

Difficultés ou pièges rencontrés :

111 lignes ont une correspondance uniquement à gauche car seulement 714 'sku' renseignées dans le dataframe 'df_web'

Jonction df_merge & df_caracteristiques (= df_merge)

Choix des attributs :

df_merge = df_merge.merge(df_caracteristiques, on = 'post_name', how = "left", indicator = True)

Clés utilisées :

'post_name' dans les 2 dataframes

Vigilances particulières au cours du traitements :

S'assurer de la correspondance des observations lors de la jonction

Difficultés ou pièges rencontrés :

214 lignes ont une correspondance uniquement à gauche car seulement 611 observations dans le dataframe 'df_caracteristiques'

Analyses univariées du prix

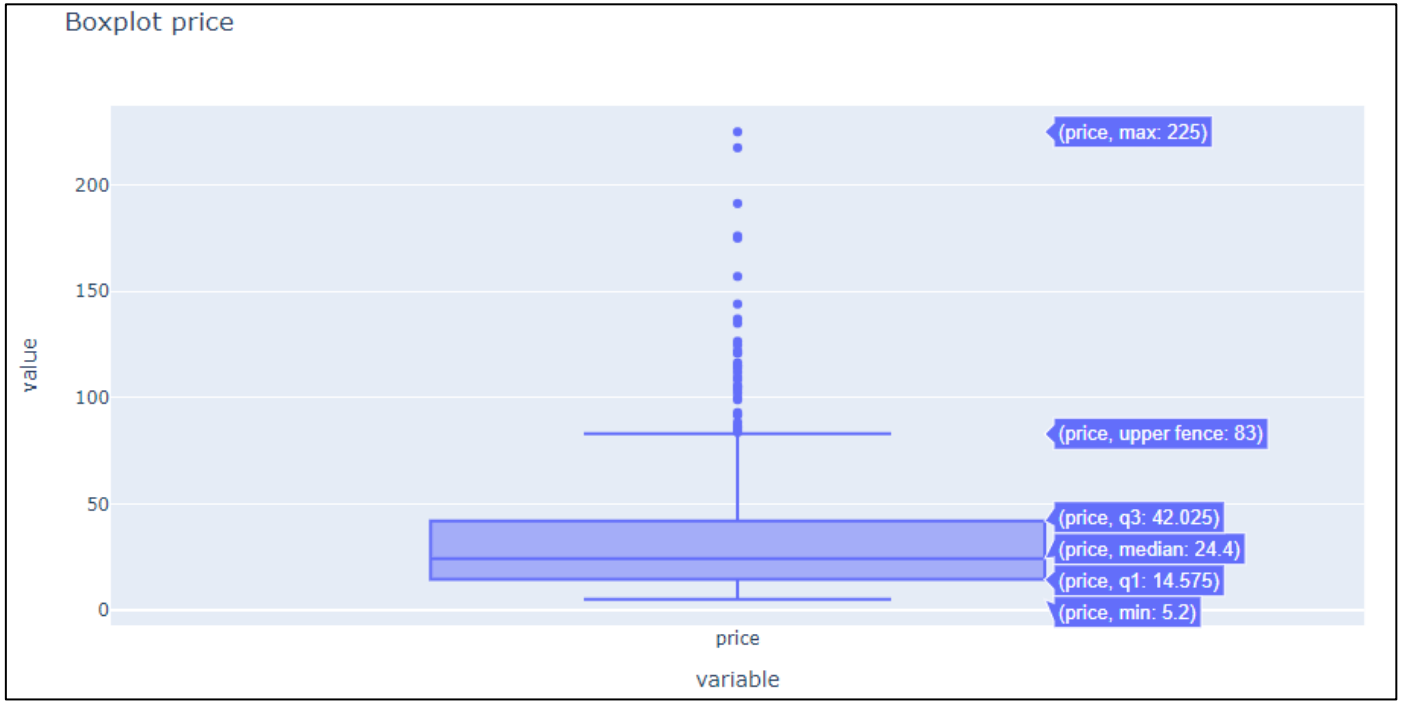
Méthodes statistiques employées :

Z-score
La moyenne de la variable prix est : 32,42€
L'écart-type du prix est de : 26,795849199710535
Le z-score est de : 1,2098888808625712
Le seuil prix pour le z-score de 3 est : 114€

Intervalle interquartile
L'intervalle interquartile est de : 27,45€
Le seuil pour les articles outliers est de 83,1€

Commentaires du graphique :

Grande disparité concernant les prix des articles vendus (min/max)



Limites éventuelles de l'analyse : Difficulté de définir les articles outliers (valeurs aberrantes [suppression] ou atypiques [conservation])avec les données brutes.

Analyses univariées du CA

Méthodes statistiques employées:

Calcul Total CA
Le chiffre d'affaire du site web est de 70568,60€

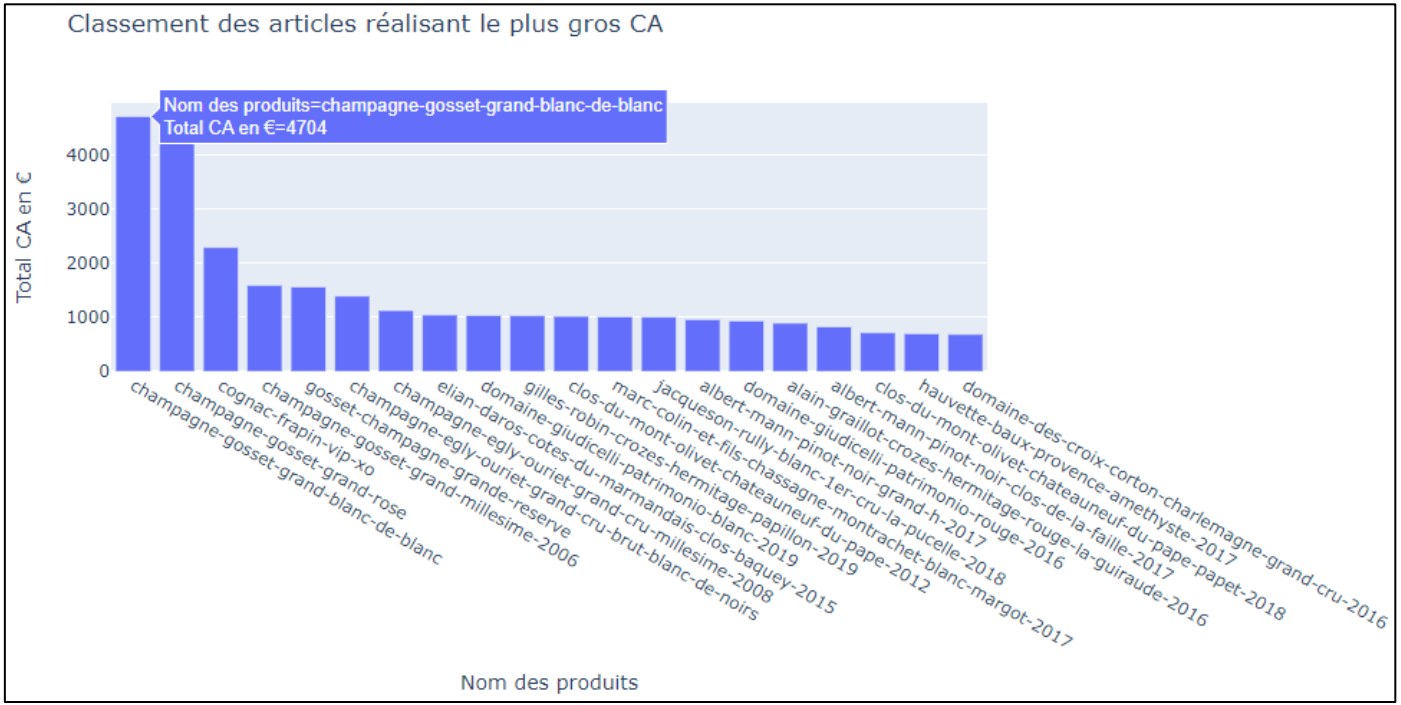
Calcul du 20/80 en CA

351 articles représentent 80% des ventes en CA
Ces articles représentent 49,16% du catalogue entier

Commentaires du graphique :

6 des 7 articles réalisant le plus de CA sont des champagne

Les Domaines Gosset et Ouriet sont les plus prolifiques en terme de CA, accompagnés du cognac Frapin



Limites éventuelles de l'analyse : Données manquantes sur certains produits pouvant potentiellement fausser l'analyse

Analyses univariées des quantités

Méthodes statistiques employées:

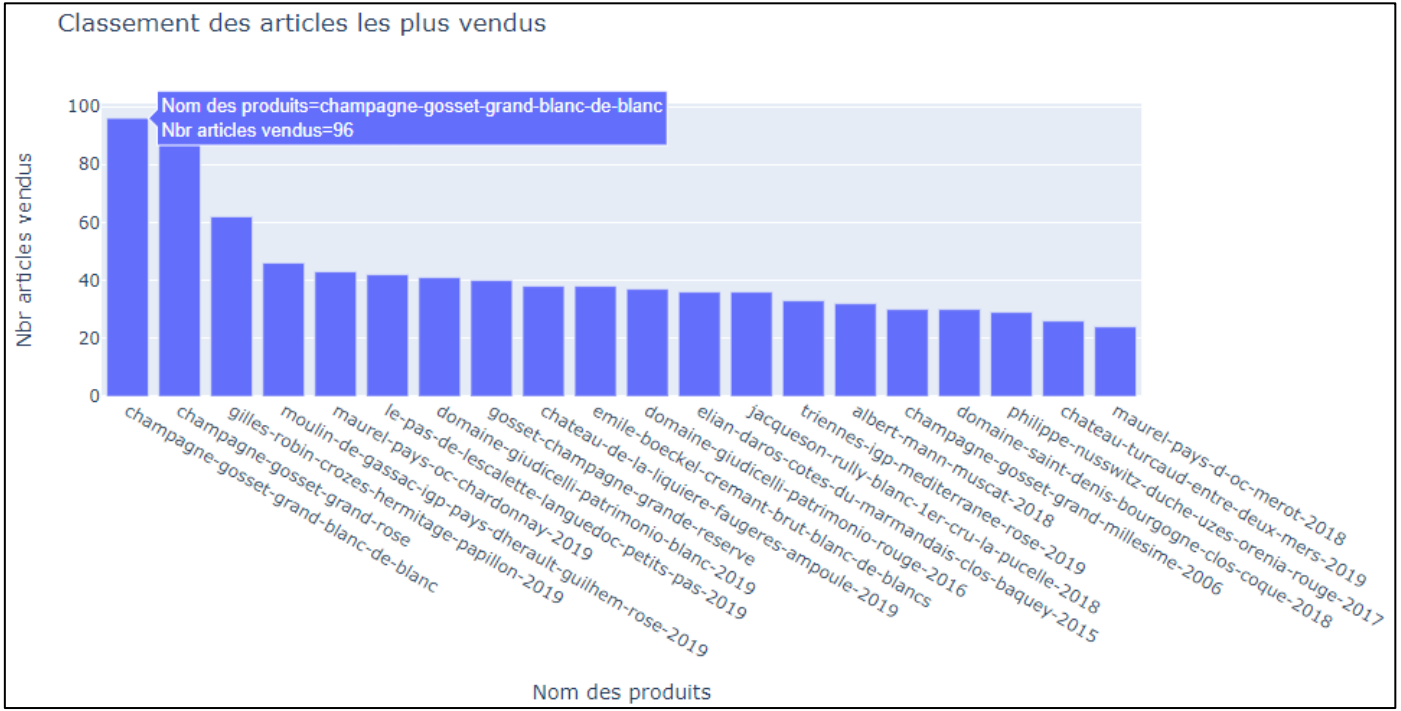
Calcul du 20/80 en quantité

366 articles représentent 80% du CA en quantité
Ces articles représentent 51,26% du catalogue entier

Commentaires du graphique :

Les 4 références champagne de Gosset font également partis des articles les plus vendus en terme de quantité

Les millésimes de 2019 représentent 40% des articles les plus vendus en quantité (8/20)



Limites éventuelles de l'analyse : Données manquantes sur certains produits pouvant potentiellement fausser l'analyse

Actions pour la suite

1 – Modifier les **‘sku’** des articles qui ne respectent pas la règle de codification (= **‘13127-1’** et **‘bon-cadeau-25-euros’**)

2 – Modifier l’**’id_web’** de l’article qui ne respecte pas la règle de codification (= **‘14680-1’**)

3 – Tenter de renseigner les données manquantes des **611 articles** ayant pour statut **‘both’** dans la colonne **‘_merge’** du fichier « **df_merge. xlsx** » en réalisant des recherches ou en effectuant des comparaisons avec les articles similaires déjà existants.

4 - Tenter de renseigner les données manquantes des articles ayant pour statut **‘left_only’** dans la colonne **‘_merge’** et filtrer en ne gardant que les articles possédant un **‘sku’ (= exclure les ‘Vides’)** dans le fichier « **df_merge. xlsx** » , soit **103 articles** puis réaliser des recherches ou effectuer des comparaisons avec les articles similaires déjà existants.

5 – Tenter d’identifier les articles ayant pour statut **‘left_only’** dans la colonne **‘_merge’** et filtrer en ne gardant que les articles ne possédant pas de **‘sku’ (= ne cocher que les ‘Vides’)** dans le fichier « **df_merge. xlsx** », soit 111 articles possédant un ‘product_id’ mais n’ayant aucunes informations concernant le produit ou les ventes associées.

Point sur les compétences apprises

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

Le travail de nettoyage s’est bien déroulé dans l’ensemble, les indications fournies par Sylvie dans le notebook étaient claires et m’ont permis de rendre un travail certes perfectible mais satisfaisant sur l’attente.

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

Les ‘choix’ ont été le plus difficile à mon sens. Mon manque d’expérience m’a rendu compliqué le fait de devoir choisir quelles variables et quelles observations étaient à conserver ou à supprimer.

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

La définition et l’application des fonctions sur les dataframes
La jonction/fusion des dataframes entre eux (choix des attributs)