ANALYSE DU STOCK ET DES VENTES DU SITE BOTTLENECK

WOUOGOU Hélène
BI-Analyst
21 Août 2023

Analyses Exploratoires des Données (fichier erp.xls)

Caractéristiques	Traitements réalisés	Remarques
825 observations et 5 colonnes(dont 3 entiers,1 décimal et 1 objet)	***Nettoyages des données:*** - Vérification de la typologie: variables importées dans le bon type - Vérification des doublons sur la colonne product_id(pas de doublons) - Valeurs manquantes: aucune -Traitement des outillés(valeurs aberrantes):comparaison variables stock_quantity et stock_statut révèle une incohérence au niveau de la variable stock ***Features engineering***: création de la colonne stock_statut_2 en tenant compte de stock_quantity ***Analyse exploratoire de chaque variable du fichier erp.xlsx*** -Analyse de la variable PRIX: les prix varient entre 5.2 et 225€ ; tous les articles ont des prix renseignés -Analyse de la variable STOCK: stocks min 0 et stocks max 578 bouteilles -Analyse de la variable ONSALE_WEB: 717 articles/825 sont en vente en ligne; Suppression des colonnes redondantes dont l'information est inutile pour la suite(utilisation de la méthode DROP de la bibliothèque panda) colonnes conservées 'stock_status', 'stock_quantity', 'resultat_diff_st_stock'	Le piège éventuel était de pouvoir bien identifier les valeurs aberrantes

Analyses Exploratoires des Données (fichier web.xlsx)

Caractéristiques	Traitements réalisés	Remarques
1513 observations et 28 colonnes(dont 3 entiers,10 decimals,11 objets,4 datetimes)	***Nettoyages des données:*** -Vérification de la typologie: variables importées dans le bon type -valeurs manquantes: plusieurs, mais 'tax_class', 'post_content', 'post_password', 'post_content_filtered' n'ont aucune valeurs renseignées -Vérification des doublons: tous les codes sku sont en doublons -Traitement des outillés(valeurs aberrantes) : La variable sku présente 4 valeurs dont les code articles ne respectent pas les règles de codification (lignes 0,787,1209,1511) ***Features engineering***: -Suppression des colonnes qui n'ont aucune valeur "virtual", "downloadable", "rating_count", "average_rating", "tax_class", "post_content", "post_password", "post_content_filtered", "post_parent", "menu_order", "post_mime_type", "comment_count", "post_status", "comment_status", "ping_status", "post_type"') -Sur la colonne sku transformation du code «bon-cadeau-25-euros » en un code à 5 caractères -Transformation du code 13127-1 en code 13127	Le piège était de bien identifier les colonnes à supprimer; Les lignes sans code articles ne sont pas renseignées dont ces valeurs ne sont pas exploitables et également bien corriger les codes qui ne respectaient pas les règles de codifications,

Analyses Exploratoires des Données (fichiers liaison.xlsx et caracs_vins_csv)

fichier liaison.xlsx

Caractéristiques	Traitements réalisés	Remarques
825 observations et 2 colonnes dont id_web objet et product_id entier	***Nettoyages des données:*** - Vérification de la typologie: variables importées dans le bon type - vérification des doublons - Valeurs manquantes: 91 valeurs sur la colonne id_web - Correction des id_web atypiques	91 articles sans correspondances

fichier caracs_vins.csv

Caractéristiques	Traitements réalisés	Remarques
611 observations et 13 colonnes (dont 12 objets et 1 float64)	***Nettoyages des données:*** - Vérification de la typologie: variables importées dans le bon type - Vérification des colonnes avec des informations manquantes: 10 variables sur 13 ont au moins une valeur manquante	Impossible de corriger ces informations manquantes ;le fichier caractéristique a un encodage spécial Windows-1252', il a fallu importer la bibliothèque chardet pour re-encoder avant de l'importer

Fusion ou consolidations des données

fichier df_erp et df_liaison

Choix des attributs	Clés utilisées	Vigilances particulières au cours du traitements	Difficultés ou pièges rencontrés
product_id	***product_id***	clé unique; df_merge_1=pd.merge(df_erp_2, df_liaison, on=cle_id, how='left')	aucune difficulté car df_erp_2 : 825 observations, df_liaison :825 observations

fichier df_merge_1 et df_web

Choix des attributs	Clés utilisées	Vigilances particulières au cours du traitements	Difficultés ou pièges rencontrés
id_web et sku	***id_web*** ***sku***	df_merge_2=pd.merge(df_merge_1, df_web_ss_vfinal, left_on=cle_id_web_1, right_on=cle_id_web_2,how='left')	2 clés différentes, df_merge_1 : 825 observations, df_web_ss_vfinal :712 observations

fichier df_merge_2 et df_caractéristiques

Choix des attributs	Clés utilisées	Vigilances particulières au cours du traitements	Difficultés ou pièges rencontrés
post_name	***post_name***	df_merge_vf=pd.merge(df_merge_2, df_caracteristiques, on=cle_post_name, how='left')	Aucune difficulté car clés communes aux 2 fichiers

Analyses univariées du prix

- *Méthodes statistiques employées:*

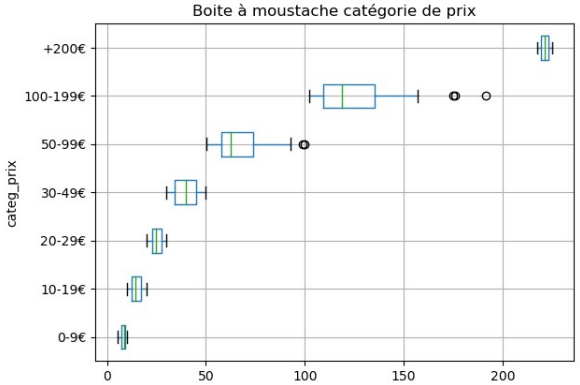
-Mesures de Tendance Centrale(moyenne, mediane)

-Mesures de Dispersion(ecart-type, variance, plage)

-Visualisations(Boîtes à Moustaches)

- *Resumé statistique et Graphique*

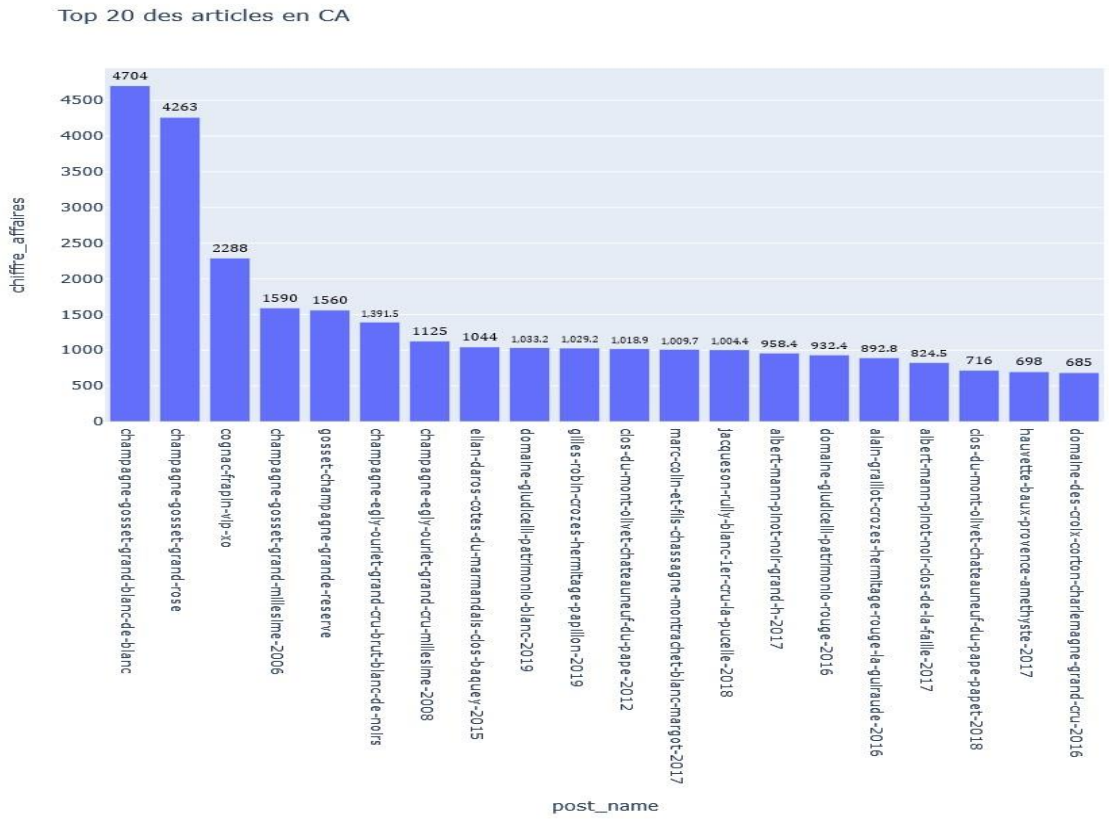
count	825.00
mean	32.41
std	26.79
min	5.20
25%	14.60
50%	24.40
75%	42.00
max	225.00



Le graphique est un box plot, il permet de visualiser la dispersion des données et identifier les valeurs extrêmes,
La boîte dans le graphique représente la plage interquartile (IQR), Les cercles ou points situés à l'extérieur des moustaches sont considérés comme des valeurs aberrantes. Nous avons 2,18% des outliers avec des prix allant de 114 à 225 €

- *Limites éventuelles de l'analyse :* **Les analyses n'ont pas été faite en fonction du temps, on aurait pu faire une analyse bivariée, l'analyse faite était purement descriptive**

Analyses univariées du CA



- Méthode statistique employée: visualisation(histogramme)
- Commentaires du graphique

Le CA des 20 premiers articles varie de 685 à 4704€,le produit avec le meilleur CA: le champagne-gosset-grand-blanc-de-blanc 4704€ suivi du champagne-gosset-grand-rose 4263 €. Les données montrent que seulement 15,76%(130/825) des produits représentent 80% du CA.

- Limites éventuelles de l'analyse : pas d'information sur la période de vente

Actions pour la suite

- Vins en stock et qui ne sont pas en vente en ligne : faudrait les mettre en ligne
- Analyser la rotation du stock car le stockage a un prix
- Automatiser la mise à jour de la base de données
- Automatiser les taches qui permettront la mise à jour des informations à chaque vente
- Analyses plus ciblées sur les vins selon les périodes(les analyses n’ont pas été faite en fonction du temps)
- Création des sites internet spécialisés pour les vins premiums

Point sur les compétences apprises

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

-Le notebook pré rempli par sylvie était un bon tremplin
-Afficher les dimensions des datasets
-Identifier les problèmes de qualité de données

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

-La densité des analyses et le manque de créativité de notre part,
-Gérez les différentes erreurs d'un jeu de données
-Comprendre la boite à moustache avec la méthode plotly express

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

-Analyse exploratoire des données