- 5 **OPTIMISEZ LA GESTION DU STOCK D'UNE BOUTIQUE EN NETTOYANT SES DONNÉES**.

CONTEXT

J'ai effectué une mission de consultation en tant qu'analyste BI **chez BottleNeck**, un célèbre vendeur de vin qui effectue des ventes en ligne. Cette entreprise souhaitait fusionner son ERP (**Enterprise Resource Planning**) avec un autre fichier contenant des informations sur les produits vendus en ligne.

L'objectif de cette conciliation était d'obtenir le chiffre d'affaires par produit, ainsi que le chiffre d'affaires global. Ils souhaitaient également savoir s'ils avaient commis des erreurs de saisie sur les prix des produits. J'ai donc dû détecter s'ils avaient des **valeurs aberrantes**, les répertorier et créer un graphique pour une meilleure lisibilité.

Méthodologie

1. Analyse exploratoire des données

- Compter, identifier les doublons, les clés uniques et les valeurs nulles ou NaN

- Identification et suppression des erreurs et des doublons

- Suppression des éléments non nécessaires à l'analyse

2. Fusion des fichiers

- Fusion des différents fichiers fournis pour pouvoir effectuer davantage d'analyses

- Choix des attributs et des méthodes (côtés) pour la fusion

- Assurer la correspondance des lignes lors de la fusion

- Exporter le jeu de données dans un fichier .XLSX pour le manager

3. Analyse univariée

- Réaliser plusieurs analyses univariées en utilisant différentes méthodes statistiques (score Z, écart interquartile, 20/80)

- Informer le client des limites possibles de l'analyse

- Créer des visualisations avec Plotly pour être plus percutant dans ma communication

- Définir les actions à entreprendre ensuite pour disposer d'un meilleur jeu de données

4. Présentation des résultats et application de la méthodologie

- Préparation d'un rapport clair et professionnel incluant la méthodologie, les résultats de l'analyse et l'évaluation des compétences

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'**
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')

Il y a deux lignes dans web qui n'ont pas de valeurs dans la colonne 'sku'. Cela signifie que ces deux produits vendus sur le site ne peuvent pas être reliés au dataframe liaison et donc au dataframe erp.
Cette absence de sku pourrait provenir d'une erreur lors de la conception du tableau Excel permettant de relier les 'product_id' et 'sku'. Il est également possible que les produits en question n'aient pas été ajoutés au niveau de l'erp.
Ces deux lignes ne trouveront pas de correspondance lors de la jointure avec df_erp_liaison, nous garderons pour la suite seulement les lignes qui ont un sku.

liaison.xlsx	**Caractéristiques :** **Nbr observations : 825 lignes Nbr variables : 2 colonnes Clés non renseignées : 9 3 Doublons détectés : 90 Clés uniques : 734**	**Traitements réalisés :** **Pas de nettoyage réalisé à ce stade**
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

Fusions ou consolidations des données

Jonction df_erp & df_liaison (= df_merge)	Jonction df_merge & df_web (= df_merge)	Jonction df_merge & df_caracteristiques (= df_merge)
Choix des attributs : merge(on='product_id', how="outer", indicator=True)	Choix des attributs : merge(df_web, left_on = 'id_web', right_on = 'sku', how = 'left', indicator = True)	**Choix des attributs :** df_merge = df_merge.merge(df_caracteristiques, on = 'post_name', how = "left", indicator = True)
Clés utilisées : 'product_id' dans les 2 dataframes	Clés utilisées : id_web' dans le dataframe 'df_merge' 'sku' dans le dataframe 'df_web'	Clés utilisées : 'post_name' dans les 2 dataframes
Vigilances particulières au cours du traitement : S'assurer de la correspondance des observations lors de la jonction.	Vigilances particulières au cours du traitements : S'assurer de la correspondance des observations lors de la jonction	**Vigilances particulières au cours du traitements :** S'assurer de la correspondance des observations lors de la jonction
Difficultés ou pièges rencontrés : Aucune difficulté rencontrée, toutes les lignes ont été correctement fusionnées.	Difficultés ou pièges rencontrés : 113 lignes ont une correspondance uniquement à gauche car seulement 714 'sku' renseignées dans le dataframe 'df_web'	

Il s'agit ici de faire une jointure, il fautre prendre en compte le fait que les product_id des ERP n'aient pas toujours une correspondance dans les sku des exports Web. On commence par une première jointure Erp/Liaison, avec product_Id en clé puisque dans les deux df ont à le même nombre de références uniques. Puis le df intermédiaire est merge à son tour avec Web_df, on utilise les sku en clé. C'est ici qu'il faut faire attention aux références non vendues sur le site.

interne (on utilise l'intersection des deux clés pour effectuer la jointure), on a le même nombre de product_id le résultat serait donc le même avec une outer joint (on utilise l'union des deux clés pour effectuer la jointure).Ensuite on fait une jointure par la droite pour rapprocher les lignes désignées par un sku tout en conservant tous les produits(la jointure est réalisée en utilisant seulement la clé de la dernière table)

Analyses univariées du prix

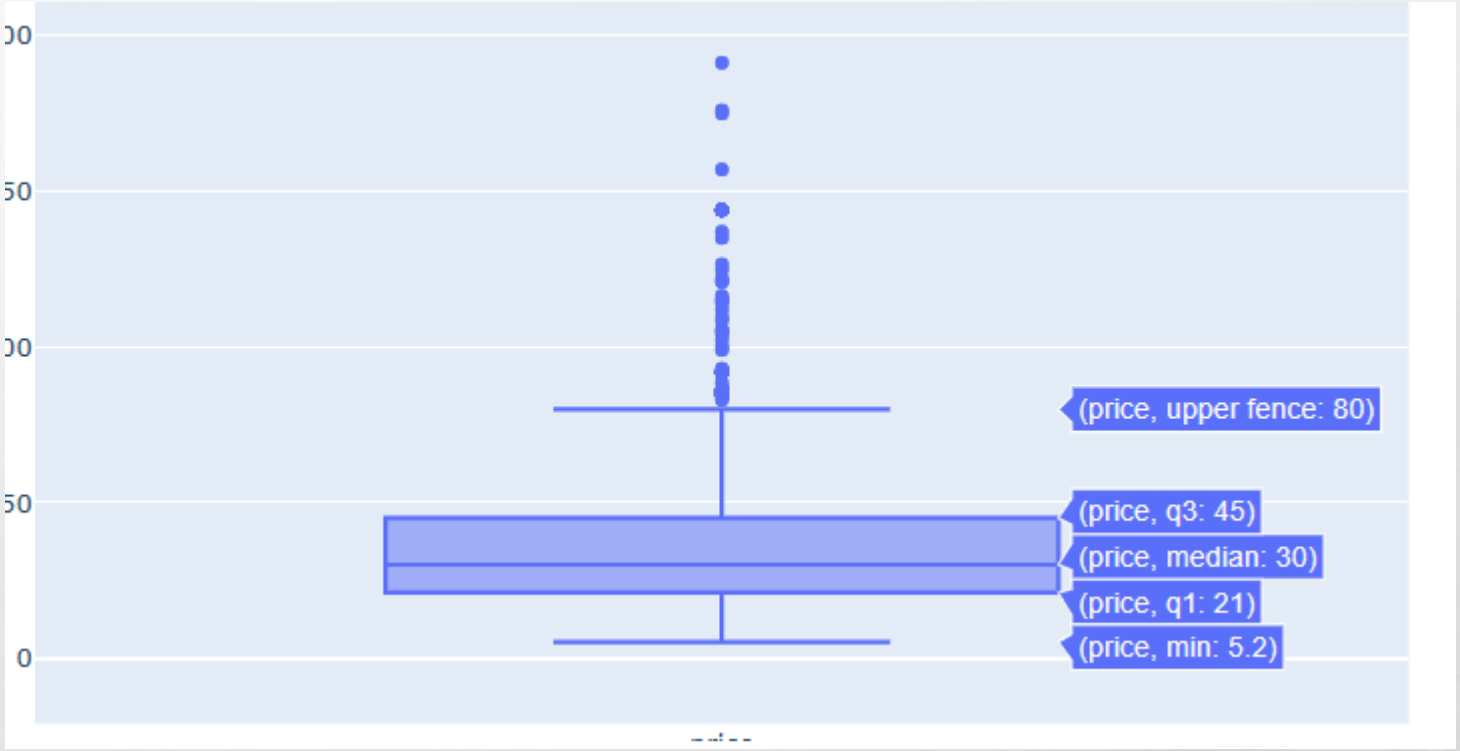
1

Méthodes statistiques employées :

Z-score La moyenne de la variable prix est : 33,42€
L'écart-type du prix est de : 21,795849199710535 Le z-score est de : 1,2098888808625712

Le seuil prix pour le z-score de 3 est : 114€

2



3

Commentaires du graphique : Grande disparité concernant les prix des articles vendus (min/max)

Sur ce graphique, les outliers sont indiqués par des cercles noirs, on observe que plusieurs valeurs entre 80 et 250 environ sont considérées comme tels. Pour afficher graphiquement les outliers, la fonction boxplot a utilisé la méthode de l'intervalle interquartile. Pour obtenir la liste de ces outliers, nous devons appliquer manuellement cette méthode.

Analyses univariées du CA

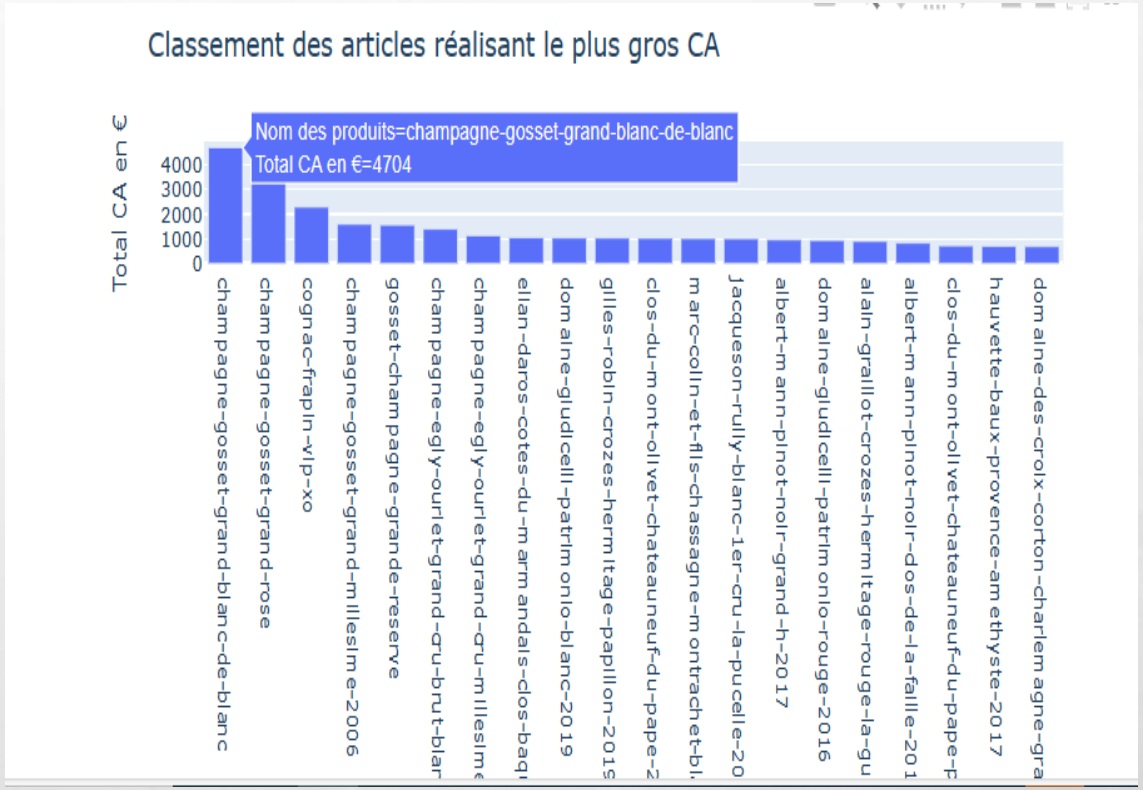
Méthodes statistiques employées:

Calcul Total CA Le chiffre d'affaire du site web est de
70318.6€

Calcul du 20/80 en CA
349

La proportion de ce nombre d'articles dans le catalogue entier est de : 49.02%

Cela signifie que parmi tous les articles vendus, seulement -349 articles (qui représentent environ 20% du total des articles) génèrent 80% du chiffre d'affaires total du site web. Cela met en évidence l'importance de ces articles dans la génération de revenus pour le site.



On retrouve nos Q1, Q3 et la médiane, les 41 outliers au-dessus de Q3 et leur position en fonction de leur prix.

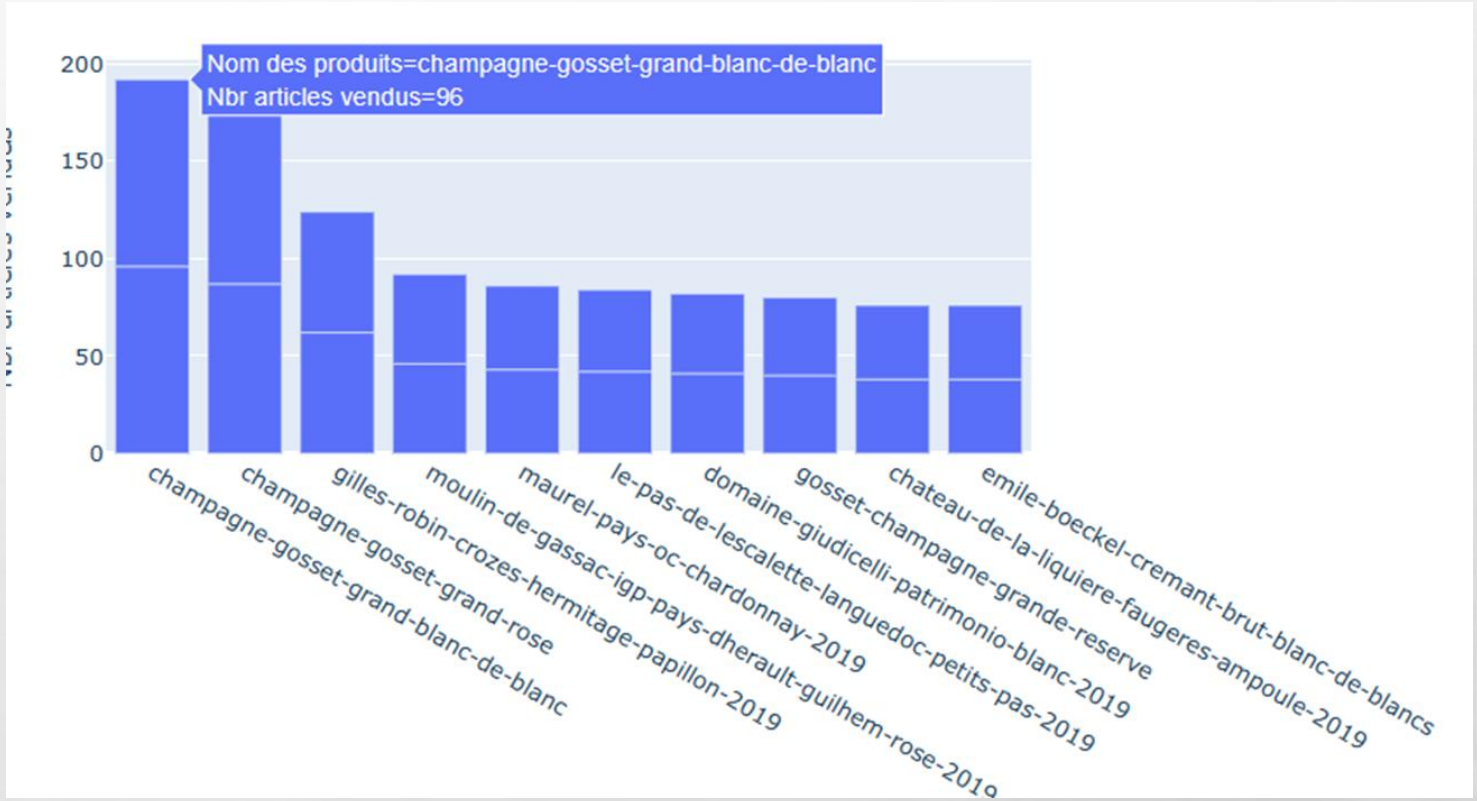
Analyses univariées des quantités

Méthodes statistiques employées:

1. Calcul du 20/80 en quantité 366 articles représentent 80% du CA en quantité

2. Ces articles représentent 51,26% du catalogue entier

3. 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



Astuce : Davantage d'options sont disponibles pour Morphose sous Options d'effet.

On remarque que les deux expressions donnent le même résulat, ce n'est dû qu'au fait que les ventes en boutiques ne sont pour le moment pas comptabilisées dans notre dataframe (total_sales toujours nulles) mais si nous complétons notre df en ajoutant ces donné Il y a deux lignes dans web qui n'ont pas de valeurs dans la colonne 'sku'. Cela signifie que ces deux produits vendus sur le site ne peuvent pas être reliés au dataframe liaison et donc au dataframe erp.␤Cette absence de sku pourrait provenir d'une erreur lors de la conception du tableau Excel permettant de relier les 'product_id' et 'sku'. Il est également possible que les produits en question n'aient pas été ajoutés au niveau de l'erp.␤Ces deux lignes ne trouveront pas de correspondance lors de la jointure avec df_erp_liaison, nous garderons pour la suite seulement les lignes qui ont un sku. Conclusion

ACTIONS POUR LA SUITE

1. Modification des 'sku' non conformes :

Identifier et mettre à jour les 'sku' des articles qui ne respectent pas la règle de codification.

2. Correction de l''id_web' non conforme:

Identifier et mettre à jour l''id_web' des articles qui ne respectent pas la règle de codification.

3. Renseignement des données manquantes pour les articles avec un statut 'both':

Comparer et compléter les données manquantes des articles qui ont un statut 'both' dans la colonne '_merge'.

4. Renseignement des données manquantes pour les articles avec un statut 'left_only' et un 'sku' non vide : Rechercher et comparer les articles similaires pour compléter les données manquantes des articles avec un statut 'left_only' et un 'sku' non vide.

1. bien passé Le travail de nettoyage ?

- Le travail de nettoyage s'est bien déroulé dans l'ensemble. Les instructions fournies dans le notebook étaient claires et m'ont aidé à réaliser un travail satisfaisant qui répondait aux attentes.

2. trouvé le plus difficile

- Le choix a été la partie la plus difficile pour moi. Mon manque d'expérience m'a rendu difficile de décider quelles variables et quelles observations garder ou supprimer.

3. tâches avoir besoin de plus d'entraînement ?

- J'ai besoin de plus d'entraînement dans la définition et l'application des fonctions sur les dataframes.
- Je dois également m'entraîner davantage sur la jonction ou fusion des dataframes entre eux, en particulier sur le choix des attributs à prendre en compte.