OPTIMISEZ LA GESTION DES DONNÉES D'UNE BOUTIQUE AVEC R OU PYTHON

Claude Olukoya

Data Analysis

Octobre 2024

Analyses Exploratoires des Données

- ***<u>Données Bottleneck</u> : 3 fichiers xlsx***

	DF_ERP	DF_WEB	DF_LIAISON
Nombre lignes :	825	714	825
Nombre colonnes :	6	18	2
Clé primaire :	product_id	sku	id_web

Analyses Exploratoires des Données

Caractéristiques du dataset :

- L’extraction de l’erp (référence produit, prix et l’état du stock)
- L’extraction de notre site web (SKU, quantités vendues, description des produits, etc.)
- Une table de liaison qui permet de lier les références entre la base de données WordPress et l’extraction de l’erp de l’entreprise.

Analyses Exploratoires des Données

- ***Traitement réalisés : Nettoyages des données*** :
  - Appréhender le nombre de colonnes, lignes, les types de datas
  - Gérer les doublons si il y en a dans le DataFrame
  - Suppression des colonnes redondantes qui n'ont que des valeurs nulles
  - Vérifier la présence des valeurs manquantes dans le DataFrame

Analyses Exploratoires des Données

- ***Features engineering :***

- Dans le dataset, il y avait des colonnes inutiles ou redondantes comme "virtual", "downloadable" etc. Je les ai supprimées pour avoir un tableau plus cohérent.

- J’ai aussi utilisé la méthode cut() pour le calcul du nombre d'articles représentant 80% du CA. Celle-ci est une technique permettant de regrouper les valeurs des variables continues dans un nombre de compartiments.

Analyses Exploratoires des Données

Remarques éventuelles, pièges ou difficultés rencontrées :

DF_ERP

* 2 lignes ***stock_status*** & ***stock_status*** ne sont pas identiques
* 3 valeurs négatives présentes dans la colonne ***prix***
* 2 valeurs négatives présentes dans la colonne ***stock_quantity***
*

DF_LIAISON

91 valeurs manquantes dans ***id_web***

3 valeurs ne suivent pas la codification dans ***id-web***

DF_WEB

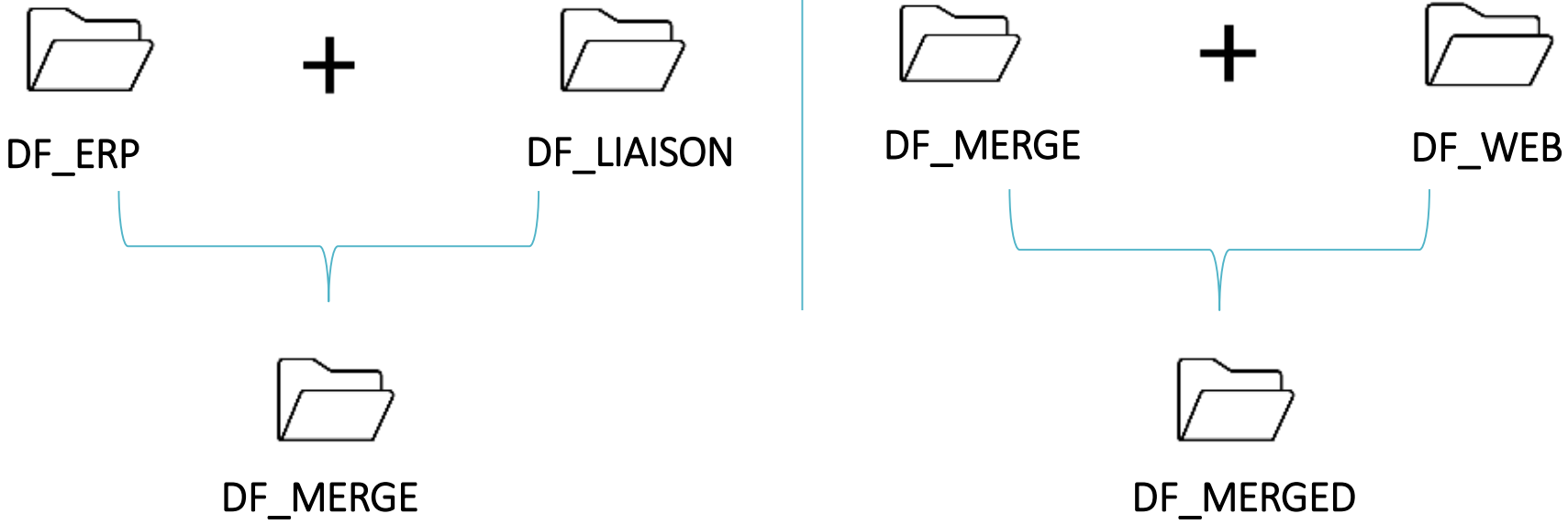
La colonne ***sku*** a des valeurs manquantes ou NaN

Il y a 2 lignes ***sku*** null avec ***total_sales*** non null

Dans ***sku***, il y a 4 valeurs qui ne suivent pas la codification

Des doublons dans post_type = ''attachment''

Fusion ou consolidations des données



Fusion ou consolidations des données

- ***Choix des attributs*** : Choisir des clés primaires communes entre deux tables pour faire une jointure

- ***Clés utilisées :*** "product_id" pour df_merge. "Id_web" et "sku" pour df_merged

- ***Vigilances particulières au cours du traitements*** :
  (i) L'utilisation de **full outer join** pour récupérer toutes les données des deux DataFrames
  (ii) Mettre "Indicator = True" comme paramètre. Celui_ci permet de voir les lignes de la jonction

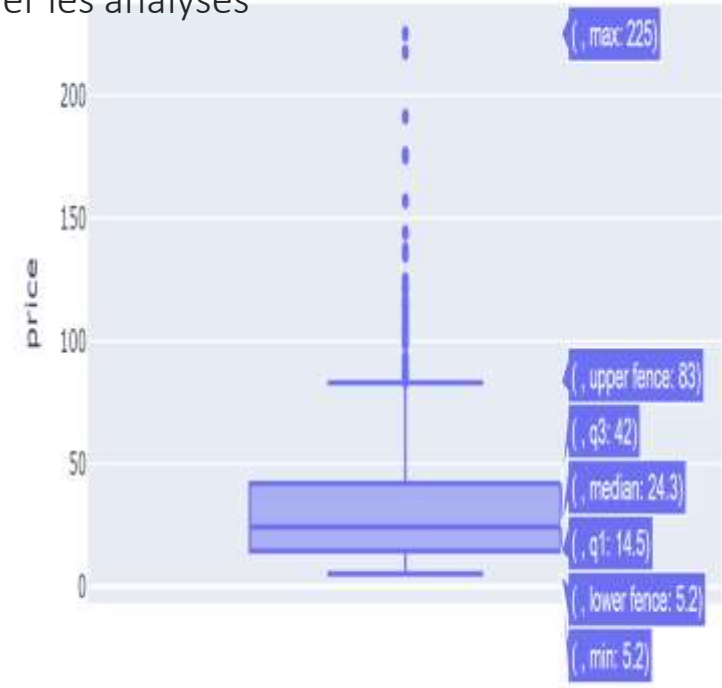
- ***Difficultés ou pièges rencontrés*** :
  (i) Dans df_merged, il y a des lignes sont pas du tout jointes
  (ii) Dans df_merge, toutes les lignes sont jointes

Analyses univariées du prix

- ***Librairie utilisée :*** Matplotlib et Plotly Express pour illustrer les analyses univariées de la variable prix.

- ***Graphique avec commentaire des résultats :***
  On peut voir directement les valeurs correspondantes
  Min: 5.2
  Q1 : 14.5
  Q2 : 24.3
  Q3 : 42
  Max: 225

- ***Limites éventuelles de l'analyse :*** Plotly express facilite le repérage des valeurs.



Analyses univariées du prix

Méthodes statistiques employés pour trouvers les outliers :

- **Z-SCORE :**

```
st.zscore(df_merged["price"])

zscore_filt = st.zscore(df_merged["price"]) > 3
z_score_outliers = df_merged.loc[zscore_filt]display(z_score_outliers.price.max())
```

- **ECART INTER-QUARTILE :**

```
Q3 = df_merged[["price"]].describe().loc["75%"]["price"]
Q1 = df_merged[["price"]].describe().loc["25%"]["price"]
Ecart_inter = Q3 - Q1print(Ecart_inter)

Valeur_faible = Q1 - (1.5 * Ecart_inter)
Valeur_elevee =  Q3 + (1.5* Ecart_inter)
```

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

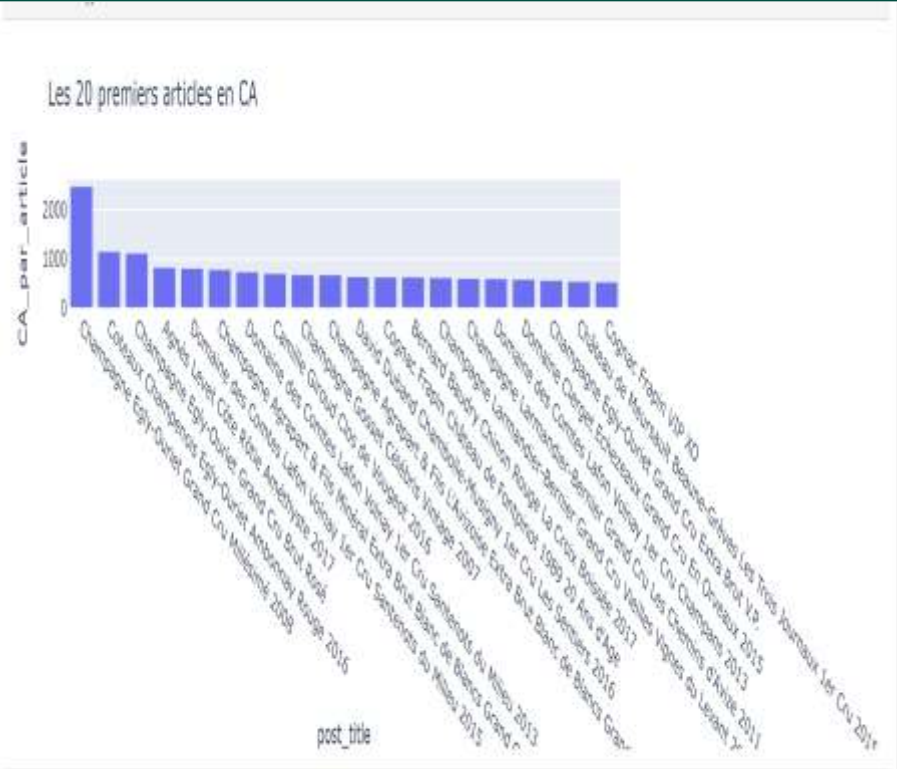
- ***Méthodes statistiques employés pour CHIFFRE D'AFFAIRES (CA) :***

  Pour trouver le CA par ligne, c'est le produit du prix et du nombre de total sales.

- ***Graphique avec commentaire des résultats :***

  Dans le graphique du top 20 premiers articles, le champagne Egly-Quriet rapporte le plus de CA et le cognac Frapin VIP rapporte le moins.

- **Limites éventuelles de l'analyse**



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

- ***Méthodes statistiques employés pour QUANTITÉS :***

  Pour trouver les vente en quantite, j'ai trié le colonne ''total sales'' pour voir le classement.

- ***Graphique avec commentaire des résultats :***

  Dans ce graphique, le produit ''Château de'' La Selve rapporte le de vente en quantité et le ''Chateau Plaisance '' rapporte le moins.

- ***Limites éventuelles de l'analyse***

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

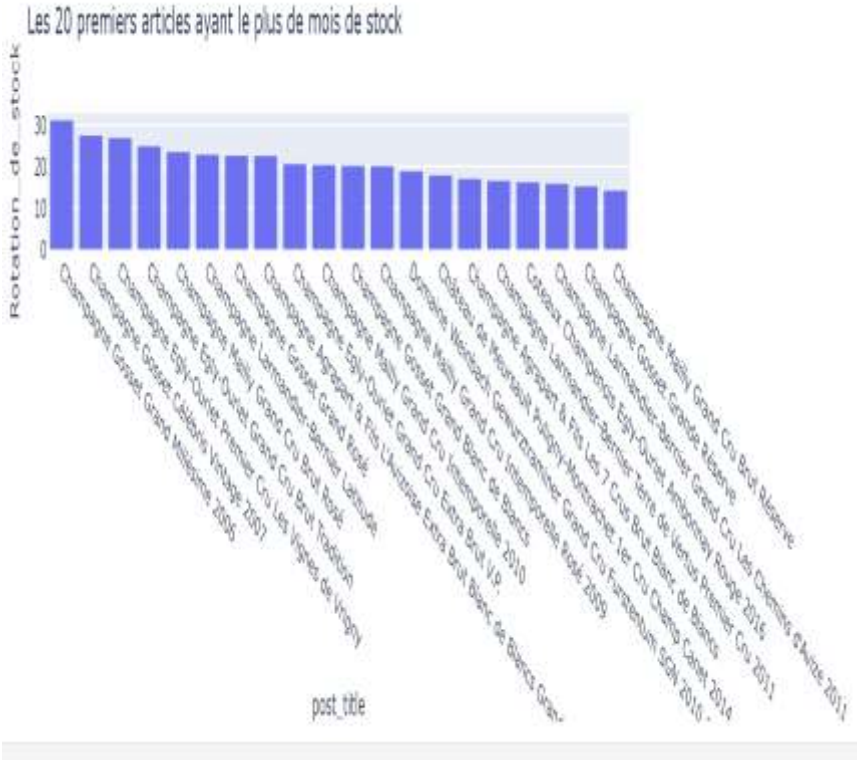
- ***Méthodes statistiques employés pour les STOCKS:***

  Pour trouver la rotation des stocks, j'ai utilisé la formule :
  (stock_quantity / total sales)

- ***Graphique avec commentaire des résultats:***

  Nous constatons ici que le Champagne Gosset est le produit qui a le plus de mois en stock tandis que le produit Champagne Mailly Grand Cru a le moins de mois de stock.

- **Limites éventuelles de l'analyse :**



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

- ***Méthodes statistiques employés pour le TAUX DE MARGE:***

  J'ai d'abord calculé le prix HT, puis j'ai calculé le taux de marge en utilisant la formule (Prix HT-Purchase price/Prix HT)

- ***Graphique avec commentaire des Résultats :***

  Le cognac a le taux de marge le plus élevé alors que le champagne a le taux de marge le plus bas.

- ***Limites éventuelles de l'analyse :***

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

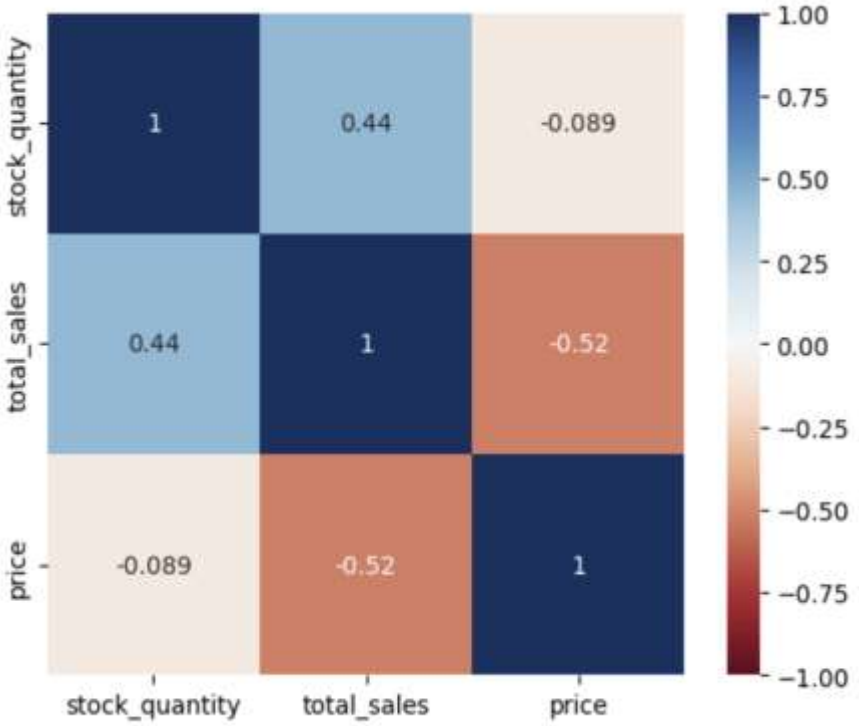
- ***Méthodes statistiques employés pour CORRÉLATIONS :***

  J'ai utilisé un heatmap Seaborn pour illustrer la corrélation entre les variables stock_quantity, total sales et price.

- ***Graphique avec commentaire des Résultats :***

  Le stock est négativement corrélé avec price car la corrélation est plutôt proche de 0 tandis que le stock est positivement corrélé avec total sales parce que la corrélation est proche de 1.

- ***Limites éventuelles de l'analyse :***



Actions pour la suite

1) Des analyses effectuées, nous pouvons tirer des conclusions de la rotation des stocks par exemple. Quels produits ont le plus de mois en stock et par la suite réapprovisionner plus tôt ces produits.

2) Les produits qui en bas du classement ‘’vente en quantité’’ doivent avoir moins de priorité.

3) Aussi, on peut analyser les ventes online pour voir si elles sont conséquentes dans le catalogue entier du CA.

Point sur les compétences apprises

- **Qu’est-ce qui s’est bien passé pour vous dans ce travail de nettoyage**?
  J’ai été très à l’aise avec le processus de nettoyage

- **Qu’est-ce que vous avez trouvé le plus difficile ?**
  Il y avait des sujets dans la partie analyse était plus compliqué car ils étaient basés sur finances, économie & gestion et il fallait faire plus des recherches pour compendre l’objectif

- **Sur quelles tâches est-ce que vous pensez avoir besoin de plus d'entraînement ?**
  J’aurai besoin de pratiquer un peu plus sur la partie analyse.