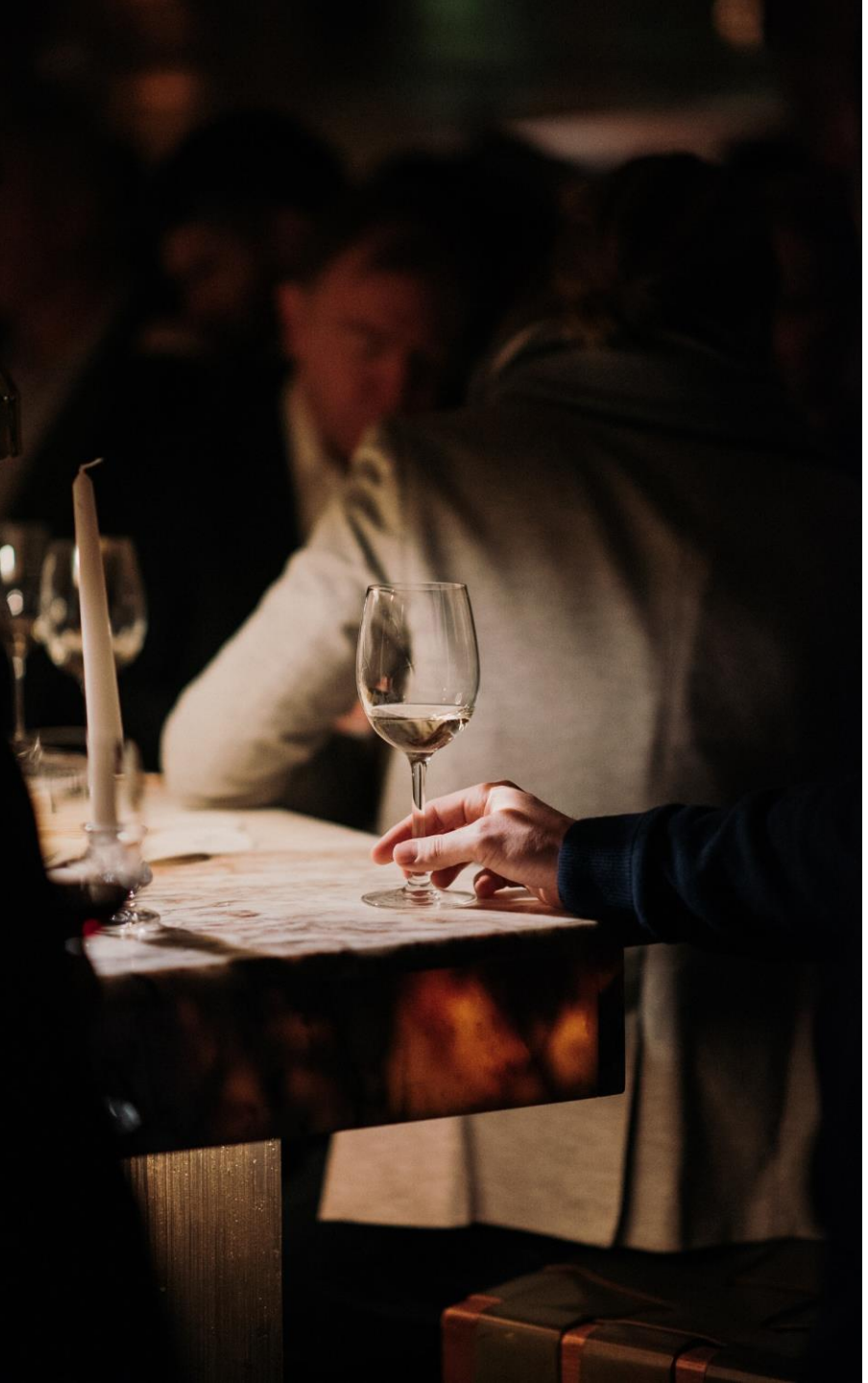
OPTIMISATION DE LA GESTION DES DONNÉES

bottleneck
fine wine spirit



Contexte

Raison pour le projet

- L'ERP n'est pas lié à notre site de vente en ligne
- On doit améliorer notre visibilité en termes d'analyse des ventes en ligne
- En attendant une solution plus centralisée, j'ai fait un rapprochement entre nos 2 bases (ERP et CMS).

Méthodologie d’analyse

3 Fichiers

- ERP – les références produit, leur prix de vente et leur état de stock
- Web - les informations des produits commercialisés en ligne
- Liaison – avec ça, nous pouvons établir le lien entre l’ID du produit dans le fichier ERP et le SKU dans le fichier Web.

Méthodologie

- Nettoyer les données - types, valeurs aberrantes, doublons, valeurs nulles, incohérences
- Tester l'unicité des clés
- Fusionner les DataFrames
- Vérifier les jointures
- Faire l’analyse

Fichier : Liaison

Mon approche

- La forme et la tête de données
- Vérifier les types de données – valeurs inattendus mais pas aberrantes.
- Somme des valeurs nulles
  - 91 valeurs nulles dans la colonne id_web.
- Comme on veut voir le chiffre d'affaires pour les ventes en ligne, j'ai supprimé toutes les lignes avec des valeurs nan dans id_web.
- Chercher des doublons – aucun
- Chercher les valeurs aberrantes dans la liste product_id - aucune.
- Changez le nom de id_web en sku.
- Tester l'unicité de la clé – 734 valeurs uniques pour product id et shape = (734, 2).

Fichier : Web

Mon approche

- La forme et la tête de données
- Les types de données - total sales et post author
- Somme des valeurs nulles
  - 4 colonnes dont toutes les valeurs sont manquantes
  - 2 lignes où il n'y a pas de SKU, mais autres colonnes ont des valeurs
  - 83 lignes avec seulement des valeurs nulles.
- Rechercher de doublons
  - Il y a des doublons pour SKU parce qu'il y a des lignes pour les post_types 'attachment' et 'product'.
- Identifier et supprimer les colonnes avec toutes les valeurs NaN ou nulles
- Supprimer les colonnes qui ne seront pas utilisées pour l'analyse
- Tester l'unicité de la clé – 714 valeurs uniques pour sku et shape = (714, 3)

Fichier : ERP

Mon approche

- La forme et la tête de données
- Les types de données
- Somme des valeurs nulles
- Rechercher de doublons
- Détection d'outliers – valeurs négatives pour price et stock quantity.
  - 2 colonnes dont les valeurs price sont négatives
  - 2 colonnes dont les valeurs stock_quantity sont négatives
- Détection des valeurs avec des incoherences
  - 1 ligne où le stock quantity est 0, mais stock_status est in stock.
  - 5 lignes ligne où le stock quantity est >0, mais stock_status est out of stock
- Tester l'unicité de la clé – 823 valeurs uniques pour product_id et shape = (823, 5)

Jointure des Dataframes

Mon approche

- Fusionner les fichiers ‘erp’ et ‘liaison’ sur Product Id avec une jointure à gauche.
- Fusionner les fichier erp-liaison et web sur SKU avec une jointure ‘inner’.
- Créer la colonne chiffre d’affaires.

Les Incoherences

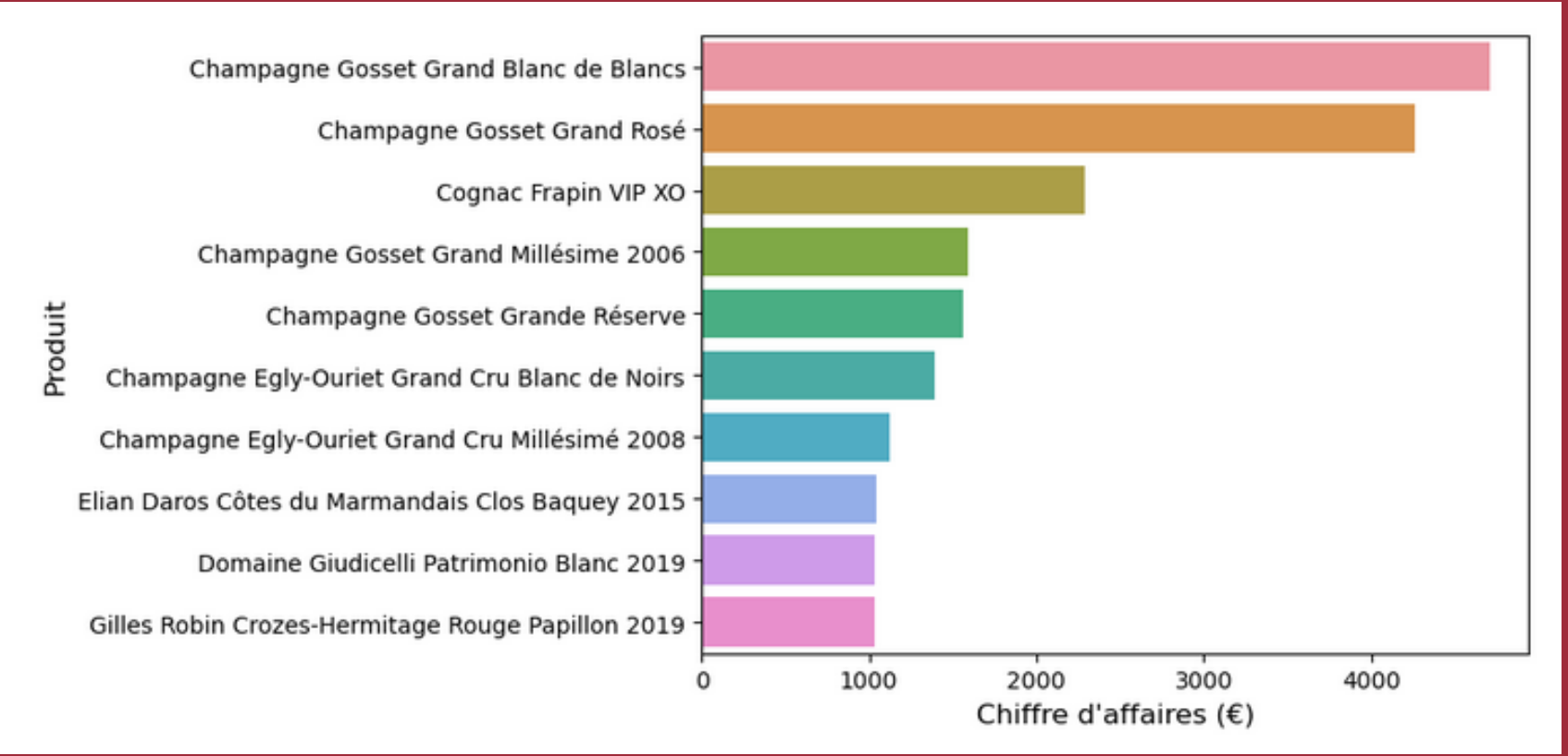
Web

- 2 lignes contiennent les valeurs NaN pour le SKU, mais elles ont des données dans leurs autres colonnes

ERP

- La colonne price a des valeurs négatives
- La colonne stock_quantity a des valeurs négatives
- Pour 1 ligne, la colonne ‘stock status’ indique instock quand le ‘stock quantity’ = 0.
- Pour 5 lignes, la colonnes ‘stock status’ indique out of stock quand le ‘stock quantity’ était > 0.

Les 10 produits avec le chiffre d'affaires le plus élevé



70568,60 €

Total du chiffre d'affaires réalisé en ligne

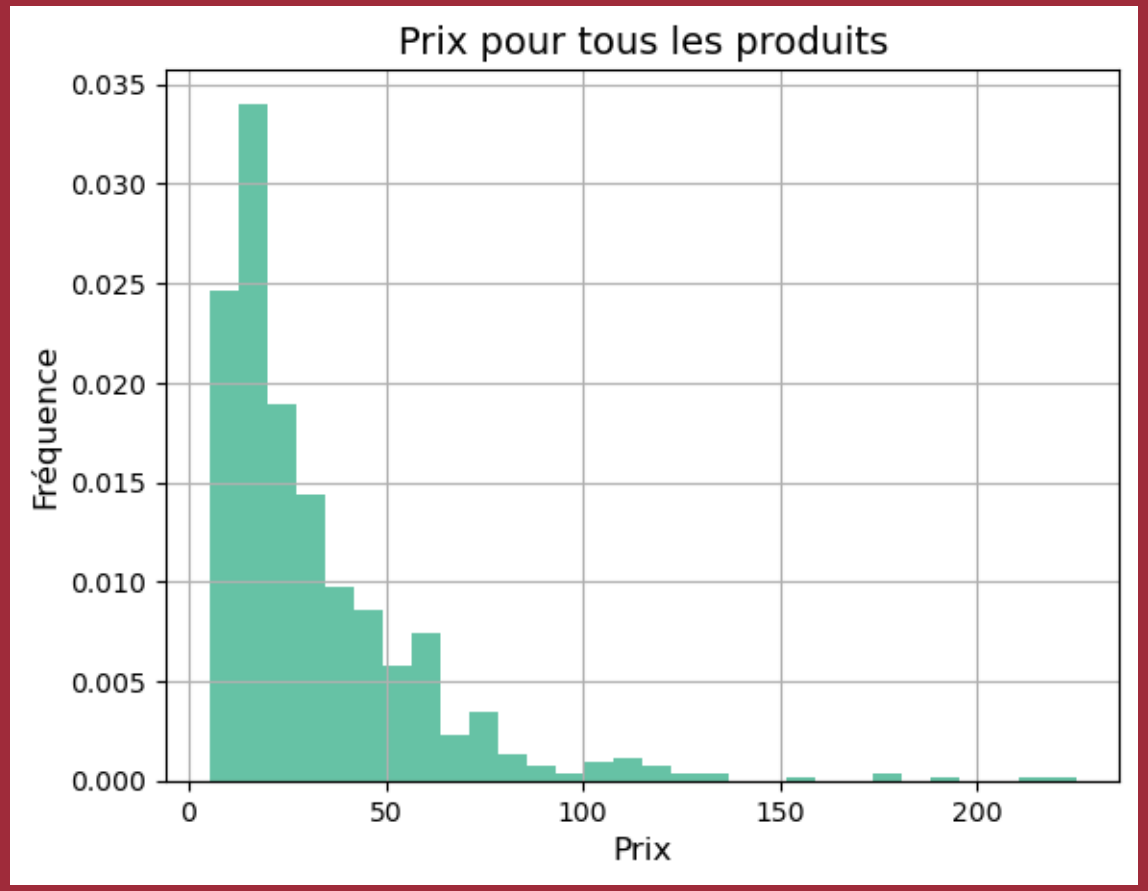
15,95 %

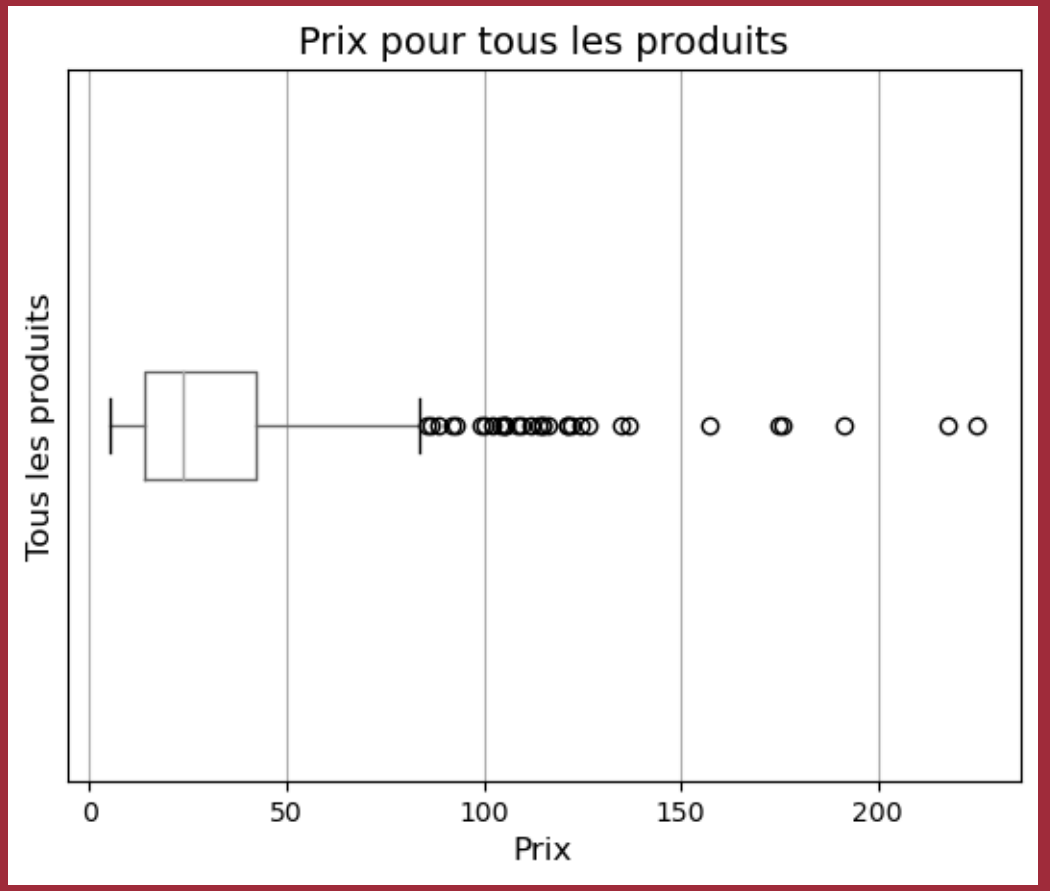
Pourcentage de notre chiffre d'affaires constitué par le top 3 produits

Analyse des prix : Chiffres

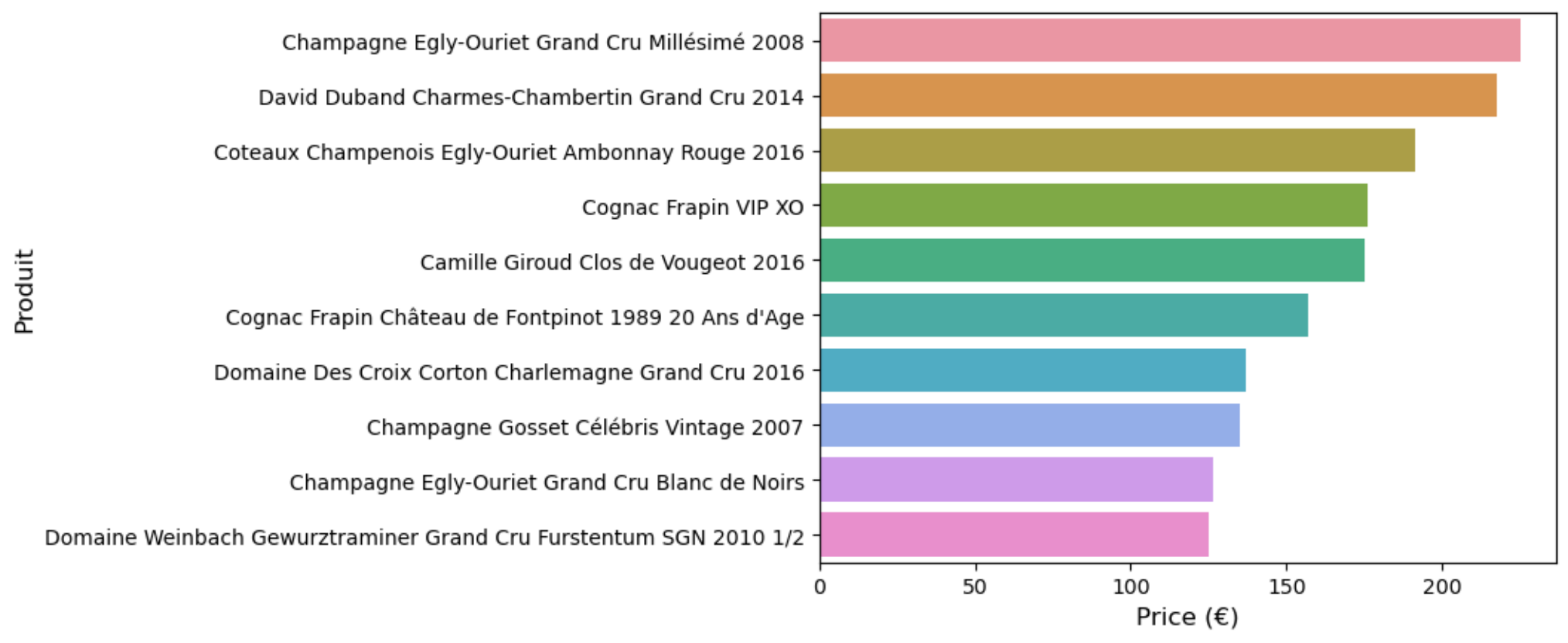
5,20 €	32,49 €	772,34	28,08
Prix minimum	Moyen prix des produits	Variance	IQR
	23,55 €	27,79	
	Médiane prix des produits	Écart type	
225 €		2,58	56
Prix maximal		Skewness	Valeurs atypiques
	19 €	10,09	
	Mode prix des produits	Kurtosis	

Analyse des prix : Graphiques





Les 10 produits avec le prix le plus élevé



Conclusion

1

Pas de valeurs aberrantes – valeurs atypiques

En excluant les valeurs négatives identifiées lors du nettoyage des données, je n'ai pas trouvé de valeurs aberrantes. Au lieu de cela, je pense que les valeurs sont tout simplement atypiques. Les prix du vin peuvent varier considérablement, et les produits les plus chers sont pour la plupart les cognacs, champagnes et grands crus, qu'on s'attendrait à être chers.

2

Un manque de données

Nous avons un certain nombre de valeurs NaN dans nos données, en particulier dans le fichier Web. Nous devrions examiner cela pour voir si nous pouvons améliorer la collecte de données.

3

Les top produits et le chiffre d'affaires

Notre chiffre d'affaires total en ligne est de 70568,60 €. Nos 3 meilleurs produits contribuent à 15,95 % de ce total. Il peut être intéressant d'étudier comment commercialiser nos produits moins populaires afin d'augmenter encore le CA.