Audit des données

Valentin - Joel - Ivan - Asma

SOMMAIRE

I. Sources dataset :

Notre principale source de données est un dataset qui prend la forme d'un fichier .csv qui proviennent d'une plateforme de commerce électronique. Il contient des informations sur les interactions des utilisateurs avec les produits.
Ces données permettent d'analyser le comportement des utilisateurs et leurs interactions avec les produits, ce qui est essentiel pour des analyses de marketing et d'expérience utilisateur.

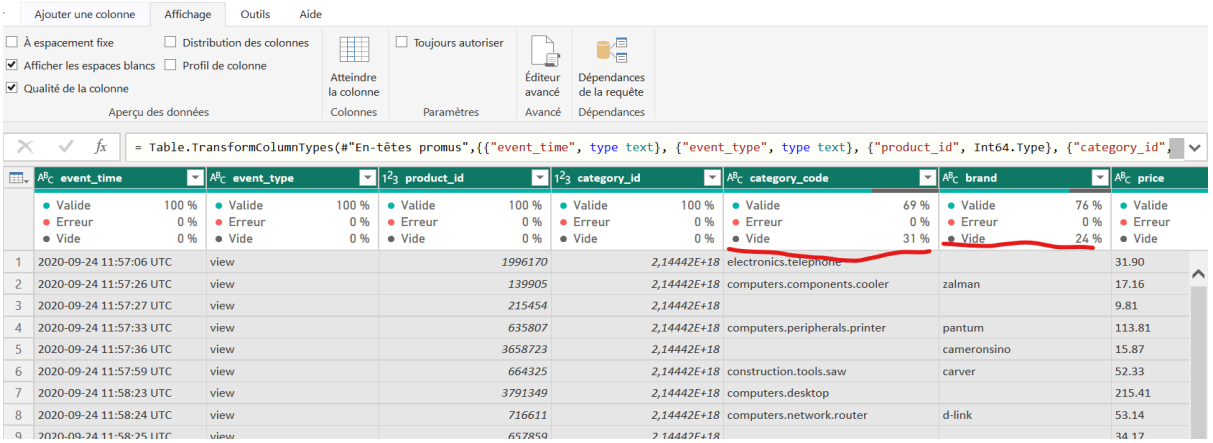
II. Types de données et description de chaque colonne :

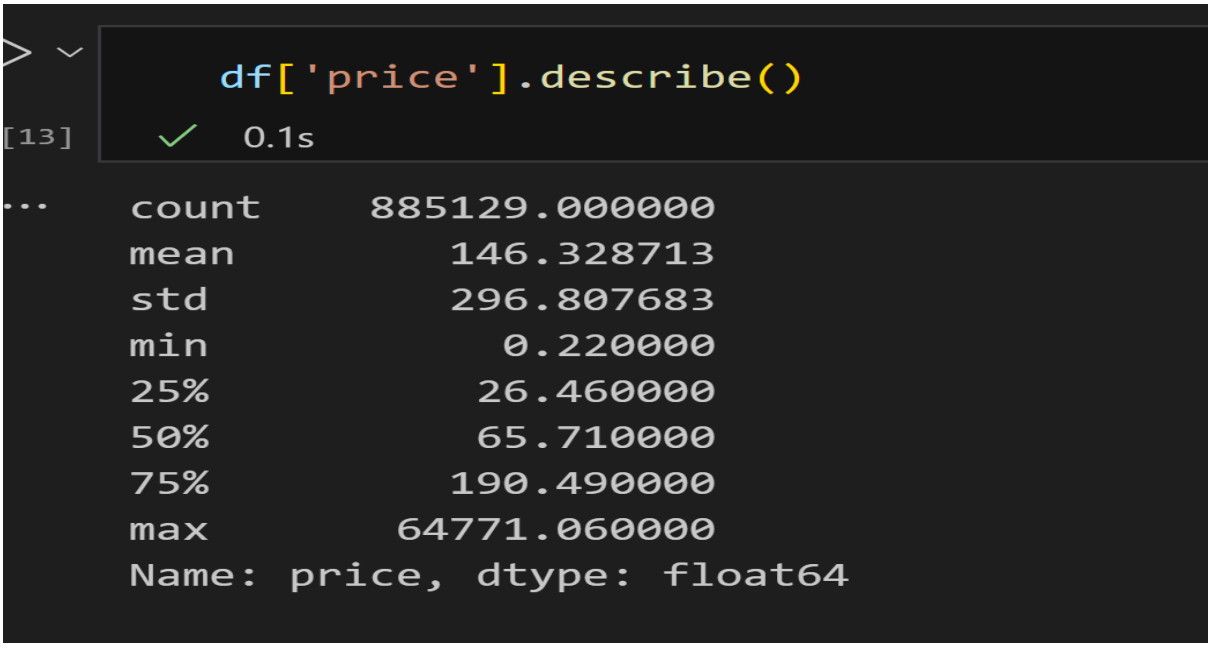
Notre dataset est composé de plus de 800 000 lignes et comprend les colonnes suivantes :

- event_time : Horodatage de l'événement.
- event_type : Type d'événement (par exemple, vue, achat, carte).
- product_id : Identifiant du produit.
- category_id : Identifiant de la catégorie.
- category_code : Code de la catégorie.(les catégories et les sous catégories sont séparé par des points)
- brand : Marque du produit.
- price : Prix du produit.
- user_id : Identifiant de l'utilisateur.
- user_session : Identifiant de la session utilisateur.

III. Qualité de chaque type de données (% de NA) :

Premiers constats: En générant le jeu de données sur Power BI, on constate au premier coup d'œil qu'il y a beaucoup de valeurs manquantes dans les colonnes brand et category_id qu'ont plus de 20% de valeur manquantes(avec 8% de valeurs manquantes dans les deux colonnes au même temps).
La colonne user_session aussi a elle-aussi quelques lignes sans valeur.on observe néanmoins une grande disparité entre la moyenne et la valeur médiane, ce qui suggère qu'il y aurait ce qui serait normalement considéré comme des valeurs aberrantes.





IV. Preprocessing des données (Découpage):

1. Suppression des doublons :
    - Nous avons identifié et supprimé les enregistrements en double. Cette étape pour garantir que chaque événement est unique, ce qui évite les biais dans notre analyse.
2. Filtrage des valeurs manquantes (NaN) :
    - Nous avons filtré et supprimé les enregistrements contenant des valeurs manquantes (NaN). Cette opération est essentielle pour éviter les problèmes lors de l'analyse et du traitement des données.

sont les deux étapes que nous avons suivies pour transformer la base de données initiale de 800,000 lignes en une base de données nettoyée de 300,000 lignes.

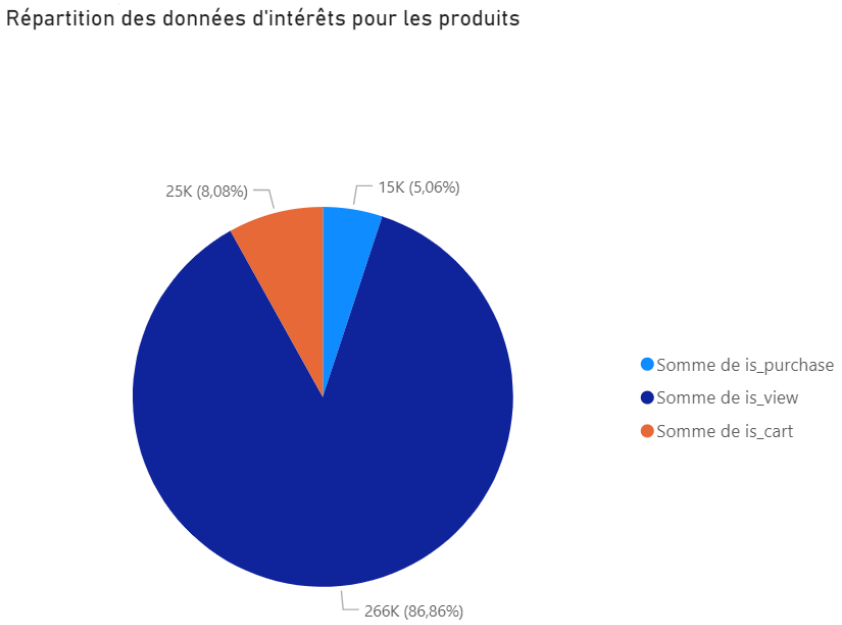
3. Transformation des string en nombres :

    - Certaines colonnes qui étaient initialement des chaînes de caractères (string), ont été transformées en valeurs numériques. Cette transformation facilite l'utilisation des algorithmes de machine learning, qui requièrent des données numériques.

4. Encodage One-Hot :

    - Nous avons appliqué l'encodage one-hot aux variables catégorielles. Cette technique crée des colonnes binaires pour chaque catégorie, permettant aux modèles de machine learning de traiter correctement ces variables sans introduire d'ordre implicite.

V. Insights des données (Graphiques présentant la répartition des données avec analyse et commentaire sur chaque colonne)



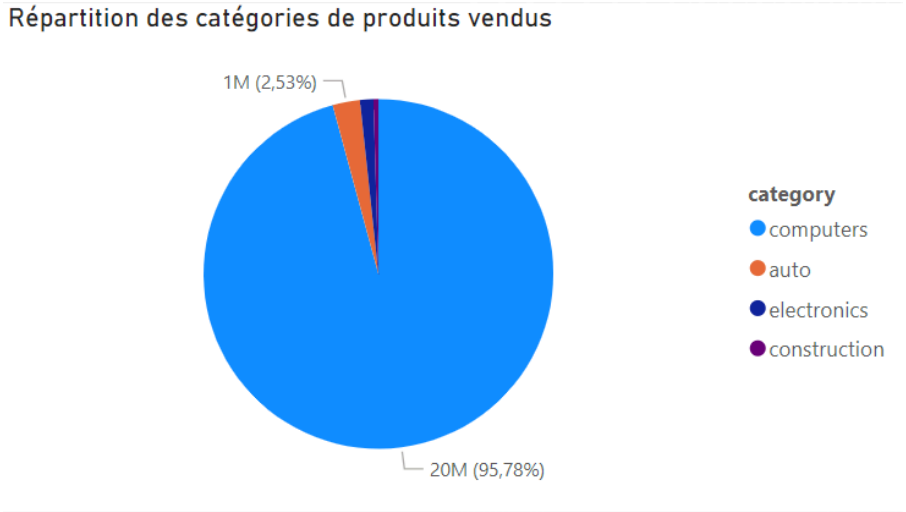
Nous avons analysé la répartition des interactions des utilisateurs avec les produits. Voici les principales catégories et leurs proportions :

1. Views (Consultations)
2. Add to Cart (Ajout au Panier)
3. Purchases (Achats)

Répartition des Interactions

- Consultations (Views) :
    - Nombre : 266,000
    - Pourcentage : 86.86%
    - Analyse : La majorité des interactions (86.86%) sont des consultations, où les utilisateurs regardent les produits sans les acheter. Cela indique un fort intérêt pour les produits, mais aussi des barrières potentielles à l'achat.
- Ajout au Panier (Add to Cart) :
    - Nombre : 25,000
    - Pourcentage : 8.08%

    - Analyse : Une portion significative des utilisateurs (8.08%) ajoutent des produits à leur panier sans finaliser l'achat. Cela suggère des intentions d'achat sérieuses, mais des obstacles dans le processus de conversion, tels que des frais de livraison élevés ou des options de paiement limitées.
- Achats (Purchases) :
    - Nombre : 15,000
    - Pourcentage : 5.06%
    - Analyse : Seulement 5.06% des interactions se traduisent par des achats réels. Ce taux de conversion relativement faible peut être attribué à divers facteurs, incluant ceux mentionnés précédemment.



Le deuxième pie chart montre la répartition des catégories de produits vendus, avec une majorité de ventes dans la catégorie "computers" suivie par "auto" et une baisse significative dans les catégories "electronics" et "construction".

Computers :

- Nombre : 20M
- Pourcentage : 95.78%

- Analyse : La catégorie "computers" domine largement les ventes, représentant près de 96% des produits vendus. Cela suggère une forte demande pour les produits informatiques.

Auto :

- Nombre : 1M
- Pourcentage : 2.53%
- Analyse : Bien que loin derrière les ordinateurs, la catégorie "auto" représente tout de même une part significative des ventes, avec environ 2.5%. Cela peut refléter une demande constante pour les produits automobiles.

Electronics :

- Analyse : La part des ventes dans la catégorie "electronics" est relativement faible. Cela suggère que, malgré la présence d'autres produits électroniques en dehors des ordinateurs, ils ne représentent qu'une petite fraction des ventes totales. Il serait intéressant d'explorer pourquoi cette catégorie est si marginale.

Construction :

- Analyse : Cette catégorie est négligeable en termes de ventes, Cela peut indiquer un marché très restreint pour les produits de construction dans ce contexte particulier, ou peut-être que les produits de construction ne sont tout simplement pas une priorité pour les clients de cette entreprise.