Audit des données

Etape 1 : L'analyse de données

Descriptif de l'ensemble de données à notre disposition

I. Sources dataset (Expliquez brièvement ce que représente le dataset)

Notre dataset contient des données du parcours client sur le site E-commerce. Elle est sous forme tabulaire contenant 885129 lignes et 9 colonnes. Notre dataset contient des doublons, des valeurs manquantes et une colonne qui doit être séparée en 3 (category_code) car elle renferme plusieurs infos dans la même colonne.

II. Description de chaque colonne

Le dataset comprend neuf colonnes répartir comme suit:

colonne	type	description
event_time	dateTime	l'heure et la date d'un événement sur un produit
event type	varchar	le type de l'événement(view, purchasse e tcate
product_id	varchar	L'id du produit
category_id	varchar	la catégorie id du produit
category_code	varchar	le code de la catégorie produit (computer, electronic, component,périphérique ..)
brand	varchar	la marque du produit (apple, samsung)
Price	numeric	le prix du produit
user_id	varchar	l'id du visiteur ou client
user_session category_code view_article_1 view_article_2 view_article_3	varchar varchar varchar varchar varchar	le login du visiteur Article consulter par les utilisateurs Article consulter par les utilisateurs Article consulter par les utilisateurs Article consulter par les utilisateurs

-

Préparation des différents preprocessing de notre data set

III. Environnement de traitement de notre donnée

Plusieurs choix d'environnement de traitement et la visualisation de la donnée s'offrent à nous notamment Excel, Locker Studio, PowerBI, Tableau, Python etc…

Pour ce projet nous avons choisi de travailler avec Python notamment Jupyter Notebook et google colab.

Chargement et visualisation de la donnée

Traitement des doublons

Une première analyse consistait à visualiser la structure de notre base de données. En effet, on a remarqué qu'il existait des doublons et procéder à leur suppression.

```
Dimension de dataset avant duplicated: (885129, 9)
Dimension de dataset après duplicated: (884474, 9)
```

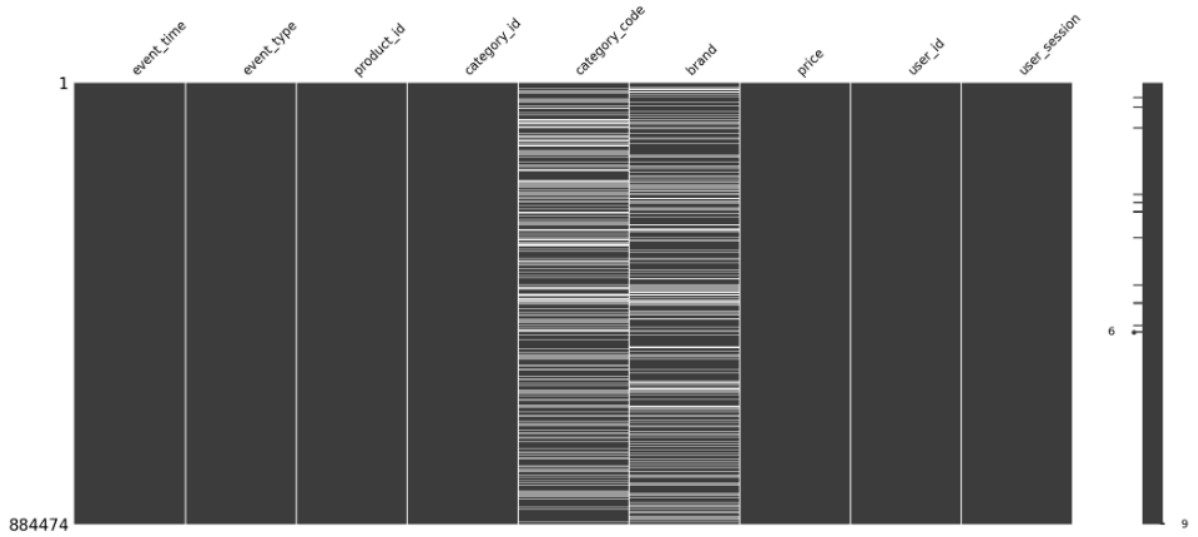
Traitement des valeurs manquantes

En statistique, on parle de valeurs manquantes lorsqu'on n'a pas d'observations pour une variable donnée pour un individu donné. Ces données ne peuvent pas être ignorées lors d'une analyse statistique. Les solutions possibles diffèrent selon la proportion et le type de ces valeurs. On pourra soit :

- Retirer les variables ou les individus présentant des données manquantes
- Imputer les valeurs

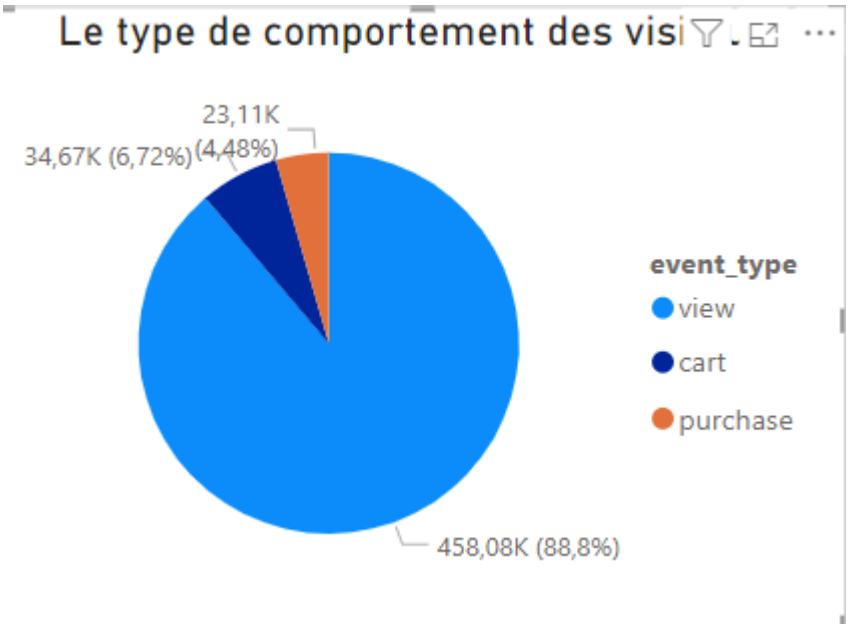
Notre jeu de données présente des valeurs manquantes sur 9 variables (voir graphe ci-dessous) avec un taux de 50.70% dont.

```
category_code    26.687522
brand            23.992435
user_session      0.018641
event_time        0.000000
event_type        0.000000
product_id        0.000000
category_id       0.000000
price             0.000000
user_id           0.000000
dtype: float64%
```

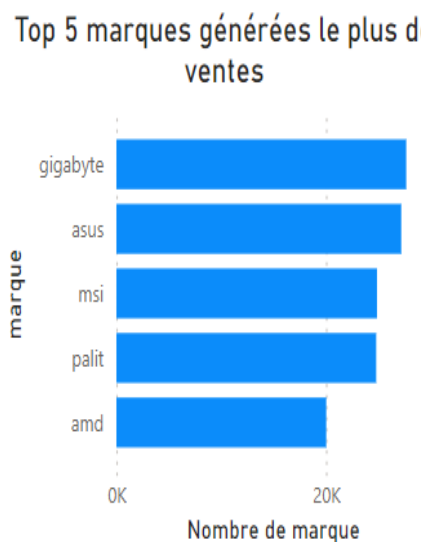


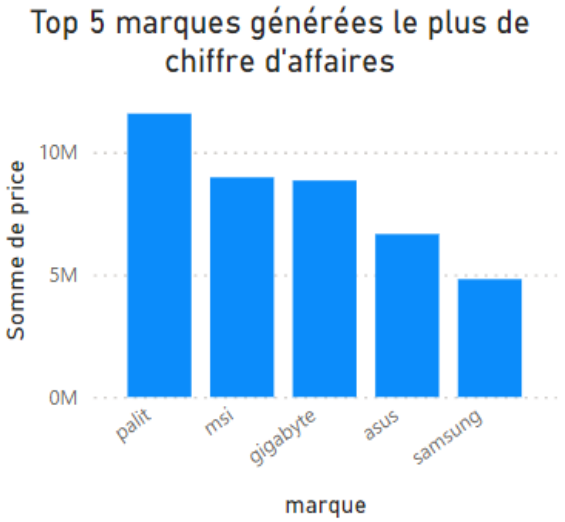
Ce graphique ci-dessus représente les données manquantes des différentes variables de notre dataset.Les bandes noires traduisent les lignes contenant des données et celles grises l'absence d'observations. Comme on peut le voir, ce sont les variables category_code, brand et user_session contenant des valeurs manquantes.

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



Ce graphique ci-dessus présente le comportement des visiteurs sur le site. On peut voir que la part des visiteurs ayant cliqué sur les produits sont plus importantes soit 88,8% que celles de ceux qui abandonnent et valide le panier.





Ces deux graphiques présentent les marques ayant générées plus de ventes et chiffres sur la période 2020-2021. On constate que les marques qui génèrent plus de ventes ne sont pas forcément celles qui génèrent le plus de chiffre d'affaires.

Ce graphique présente les 5 produits les plus vendus, on constate que sont les videcards et telephone qui sont les produits les plus ve