

Royaume de Belgique

Projet de spécialisation IA

**Rapport de nettoyage des données**

Par Harold Rajaonarison

Année Académique 2024 - 2025

**Table des matières**

[**Liste de figures** i](#_Toc199675333)

[**Introduction** 1](#_Toc199675334)

[**1.** **Chargement initial et exploration des données** 2](#_Toc199675335)

[**1.1.** **Téléchargement et extraction des fichiers sources** 2](#_Toc199675336)

[**1.2.** **Contraintes liées au volume et à la mémoire** 2](#_Toc199675337)

[**1.3.** **Lecture par blocs avec pandas** 3](#_Toc199675338)

[**1.4.** **Exploration du premier bloc de données** 4](#_Toc199675339)

[**1.4.1.** **Aperçu structurel des données** 4](#_Toc199675340)

[**1.4.2.** **Analyse descriptive du prix** 5](#_Toc199675341)

[**1.4.3.** **Répartition des types d’événements** 5](#_Toc199675342)

[**1.4.4.** **Distribution des prix** 6](#_Toc199675343)

[**1.4.5.** **Présence de valeurs nulles** 7](#_Toc199675344)

[**2.** **Stratégie de nettoyage** 7](#_Toc199675345)

[**2.1.** **Traitement des valeurs manquantes** 7](#_Toc199675346)

[**2.2.** **Formatage et types de données** 7](#_Toc199675347)

[**2.3.** **Nettoyage des valeurs aberrantes** 7](#_Toc199675348)

[**2.4.** **Filtrage des événements non pertinents** 8](#_Toc199675349)

[**2.5.** **Séparation des catégories produits** 8](#_Toc199675350)

[**2.6.** **Finalisation du jeu de données** 8](#_Toc199675351)

[**2.7.** **Vérification après nettoyage** 8](#_Toc199675352)

[**2.8.** **Sauvegarde des données nettoyées** 9](#_Toc199675353)

[**3.** **Construction du jeu utilisateur** 10](#_Toc199675354)

[**3.1.** **Objectif de l’agrégation** 10](#_Toc199675355)

[**3.2.** **Approche progressive et sécurisée** 10](#_Toc199675356)

[**3.3.** **Caractéristiques extraites** 10](#_Toc199675357)

[**3.4.** **Architecture de traitement** 10](#_Toc199675358)

[**3.5.** **Fusion finale des utilisateurs multi-chunks** 11](#_Toc199675359)

[**4.** **Analyse exploratoire** 12](#_Toc199675360)

[**4.1.** **Objectifs de l’exploration** 12](#_Toc199675361)

[**4.2.** **Statistiques générales des utilisateurs** 12](#_Toc199675362)

[**4.3.** **Analyse des comportements d’interaction** 13](#_Toc199675363)

[**4.4.** **Variables d’engagement et récence** 15](#_Toc199675364)

[**4.5.** **Préférences de produits** 16](#_Toc199675365)

[**4.6.** **Corrélations entre variables numériques** 17](#_Toc199675366)

[**Conclusion** 18](#_Toc199675367)

**Liste de figures**

[Figure 1 - Total de données avant traitement 2](#_Toc199676301)

[Figure 2 - Utilisation initiale de la RAM 3](#_Toc199676302)

[Figure 3 - Utilisation de la RAM pendant un chargement de données 3](#_Toc199676303)

[Figure 4 - Aperçu structurel des données 4](#_Toc199676304)

[Figure 5 - Analyse descriptive du prix 5](#_Toc199676305)

[Figure 6 - Répartition des types d'événements 6](#_Toc199676306)

[Figure 7 - Distribution des prix 6](#_Toc199676307)

[Figure 8 - Présence de valeurs nulles 7](#_Toc199676308)

[Figure 9 - Vérification des données nettoyées 9](#_Toc199676309)

[Figure 10 - Architecture de traitement des données en batch 11](#_Toc199676310)

[Figure 11 - Résultat de la fusion des données en un seul dataframe 11](#_Toc199676311)

[Figure 12 - Statistiques générales des utilisateurs 12](#_Toc199676312)

[Figure 13 - Distribution du nombre de vues par utilisateur 13](#_Toc199676313)

[Figure 14 - Distribution du nombre d'ajouts au panier par utilisateur 13](#_Toc199676314)

[Figure 15 - Distribution du nombre d'achats par utilisateurs 14](#_Toc199676315)

[Figure 16 - Répartition des dépenses totales par utilisateur 14](#_Toc199676316)

[Figure 17 - Nombre de jours actifs par utilisateur 15](#_Toc199676317)

[Figure 18 - Récence par utilisateur 15](#_Toc199676318)

[Figure 19 - Top 10 des catégories principales préférées 16](#_Toc199676319)

[Figure 20 - Matrice de corrélation entre les variables numériques 17](#_Toc199676320)

**Introduction**

Dans le cadre de ce projet, nous nous intéressons au comportement des utilisateurs sur la plateforme de commerce en ligne Amazing. Face à une baisse du chiffre d’affaires observée sur ses produits propriétaires, l’entreprise cherche à mieux comprendre les dynamiques d’usage et à identifier des profils clients types en vue d’améliorer son positionnement et ses stratégies de ciblage.

Les données mises à disposition couvrent sept mois d’activité et représentent plus de 50 Go de logs d’événements, incluant les vues de produits, les ajouts au panier, les achats et d'autres interactions. Un tel volume exige une approche méthodique et scalable pour le traitement, le nettoyage et l’exploration.

Ce rapport se concentre exclusivement sur la **préparation du jeu de données utilisateur**. Il documente les étapes clés de nettoyage par blocs, de transformation des événements en profils agrégés par utilisateur, et d’analyse exploratoire des comportements. Des visualisations statistiques ont été produites pour identifier les tendances dominantes, les comportements extrêmes et les déséquilibres structurels.

Les résultats obtenus révèlent une base utilisateur fortement déséquilibrée, avec une majorité de profils passifs et une minorité très active. Cette hétérogénéité justifie le recours futur à des techniques de segmentation. Le jeu final, structuré et optimisé, est désormais prêt à être utilisé pour une modélisation dédiée, qui fera l’objet d’un second livrable.

1. **Chargement initial et exploration des données**
   1. **Téléchargement et extraction des fichiers sources**

Les données nécessaires à l’analyse ont été mises à disposition via plusieurs archives compressées accessibles depuis Google Drive. Chaque fichier contenait un fichier CSV, chacun représentant un mois d’activité de la plateforme Amazing, d’octobre 2019 à avril 2020. Après téléchargement, les fichiers ont été extraits localement dans un répertoire de travail.

|  |  |
| --- | --- |
| **Nom du fichier** | **Mois concerné** |
| 2019-Oct.csv | Octobre 2019 |
| 2019-Nov.csv | Novembre 2019 |
| 2019-Dec.csv | Décembre 2019 |
| 2020-Jan.csv | Janvier 2020 |
| 2020-Feb.csv | Février 2020 |
| 2020-Mar.csv | Mars 2020 |
| 2020-Apr.csv | Avril 2020 |

Figure 1 - Total de données avant traitement

Le volume total des fichiers combinés atteint environ 52Gb soit plusieurs millions de lignes de données, ce qui nécessite une attention particulière pour la lecture et le traitement.

* 1. **Contraintes liées au volume et à la mémoire**

Le chargement direct de totalité des données via un appel unique aurait conduit à une saturation de la mémoire vive, voire à un gel du système. Une expérimentation a été réalisée à ce sujet et une hausse significative de l’utilisation de la mémoire a été observée. Ce type de jeu de données, de taille industrielle, n’est donc pas directement exploitable en mémoire sur des machines standards.

Il est donc nécessaire de maîtriser la consommation de ressources dès les premières étapes, ce qui implique une stratégie de lecture adaptée. Cette maîtrise est également indispensable pour les étapes de traitement à venir.

Voici un aperçu de l’utilisation de la RAM avant la lecture d’un fichier d’un mois.

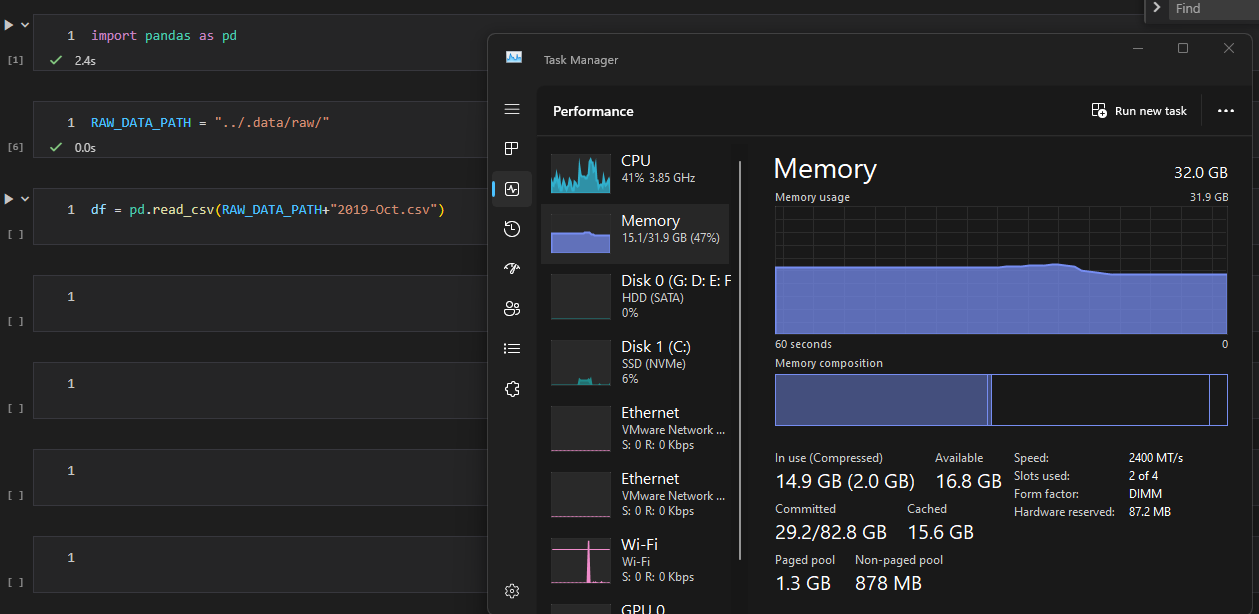


Figure 2 - Utilisation initiale de la RAM

Et ci-dessous son utilisation pendant la lecture.

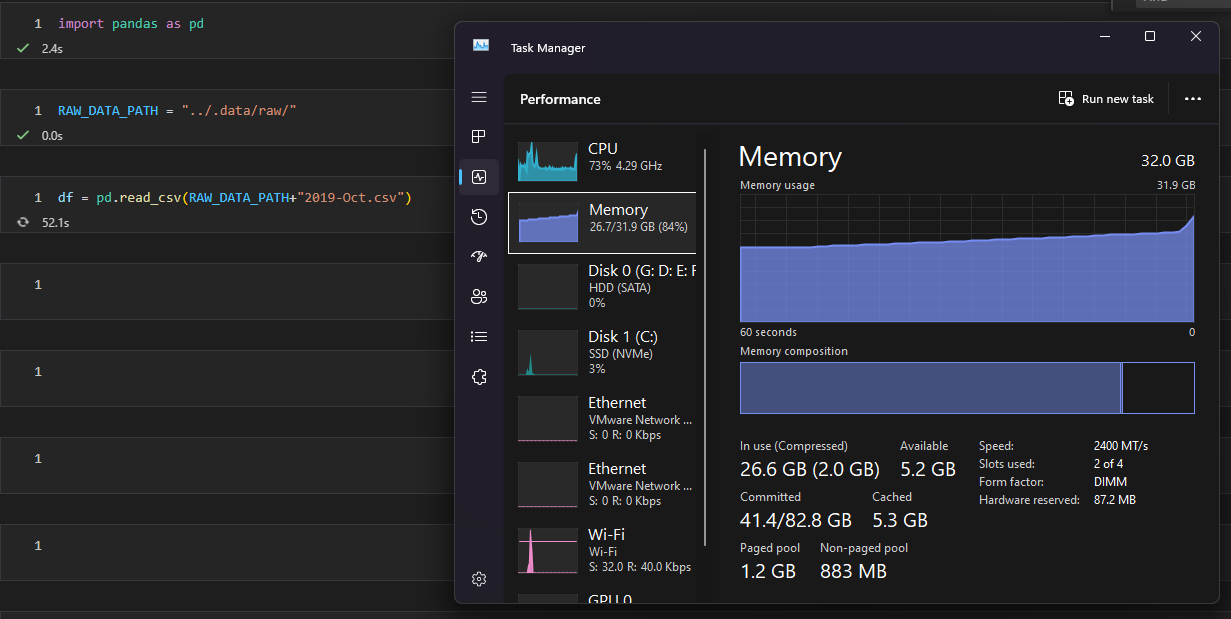


Figure 3 - Utilisation de la RAM pendant un chargement de données

* 1. **Lecture par blocs avec pandas**

Pour contourner cette contrainte, une lecture progressive en blocs a été mise en œuvre. Cette fonctionnalité permet de lire un fichier par segments successifs, sans en charger la totalité en mémoire.

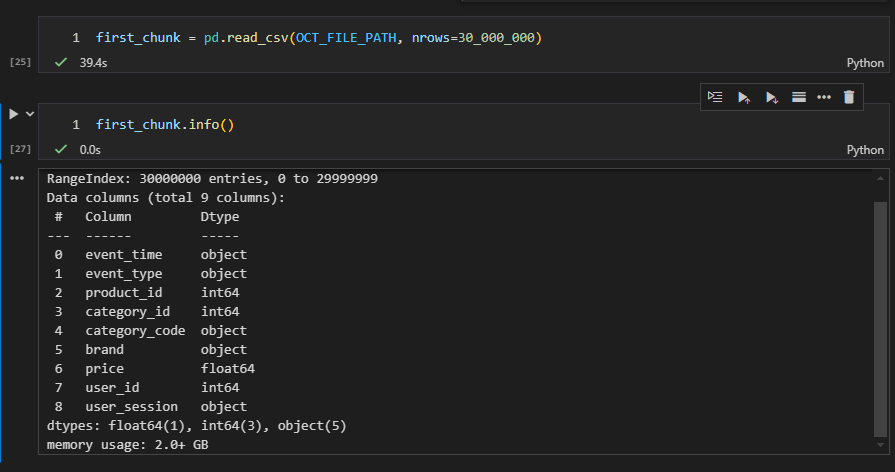
Un seuil de 30.000.000 de lignes par bloc a été défini, suite à plusieurs expérimentations par rapport aux ressources à disposition, ce qui permet un bon compromis entre performance d’extraction et stabilité de l’environnement. Chaque bloc est traité séparément selon une logique de pré-nettoyage.

* 1. **Exploration du premier bloc de données**

Avant d’appliquer le traitement sur l’ensemble des fichiers ou blocs, une analyse exploratoire préliminaire a été réalisée sur le premier chunk issu du fichier 2019-Oct.csv. Cette étape est cruciale pour comprendre la structure et la qualité des données, et permet d’identifier les opérations de nettoyage nécessaires de manière éclairée.

* + 1. **Aperçu structurel des données**

Le premier bloc de données analysé contient 30.000.000 d’entrées, correspondant aux premiers événements du mois d’octobre 2019.



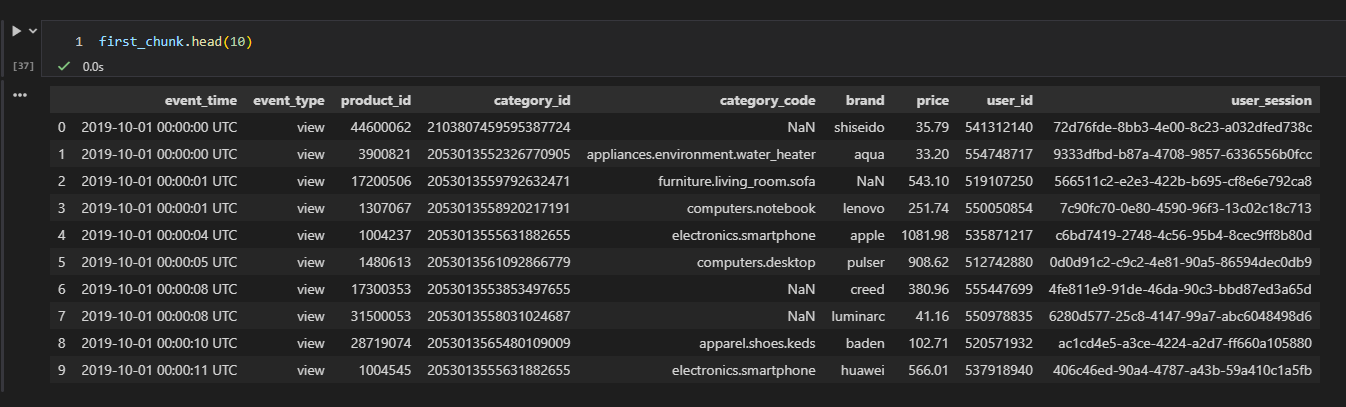


Figure 4 - Aperçu structurel des données

On observe que le fichier se compose de 9 colonnes principales représentant les interactions des utilisateurs sur la plateforme : informations temporelles, actions effectuées, identifiants des produits, métadonnées produits et données utilisateur.

* + 1. **Analyse descriptive du prix**

L’analyse des prix montre une grande variété de montants, allant de 0 € à plus de 2 500 €. Le prix moyen est d’environ 290 €, mais la plupart des produits sont en réalité bien moins chers, comme le montre la médiane à 160 €.

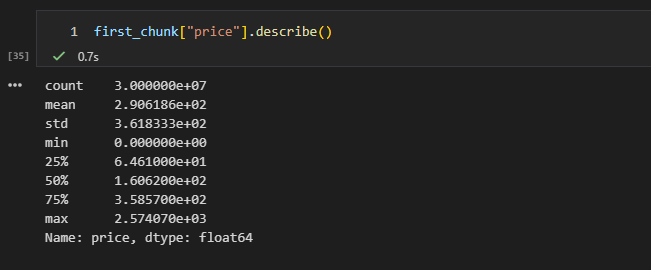
Certaines valeurs sont incohérentes et devront être écartées lors du nettoyage.

Figure 5 - Analyse descriptive du prix

* + 1. **Répartition des types d’événements**

Le graphique ci-dessous montre la fréquence des différentes actions effectuées par les utilisateurs. On observe que la grande majorité des interactions concernent la simple consultation de produits (view), tandis que les ajouts au panier (cart) et les achats (purchase) sont beaucoup moins fréquents. Cela reflète un comportement habituel sur les plateformes de vente en ligne, où les utilisateurs explorent de nombreux articles avant de passer à l’acte d’achat.

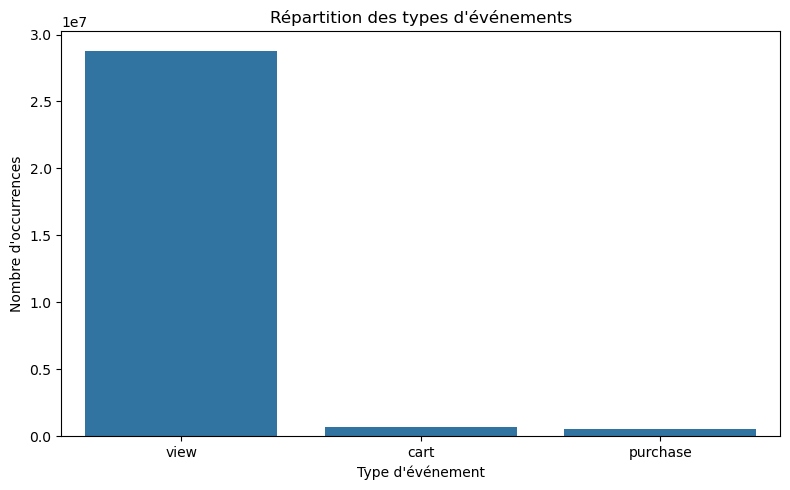


Figure 6 - Répartition des types d'événements

* + 1. **Distribution des prix**

Le graphique ci-dessous montre la répartition des prix des produits consultés ou achetés par les utilisateurs. La majorité des événements concernent des produits à moins de 100 €, avec une forte concentration entre 0 € et 200 €. Les produits plus chers sont plus rares. Cette distribution confirme que la plupart des clients s’intéressent à des articles relativement abordables.

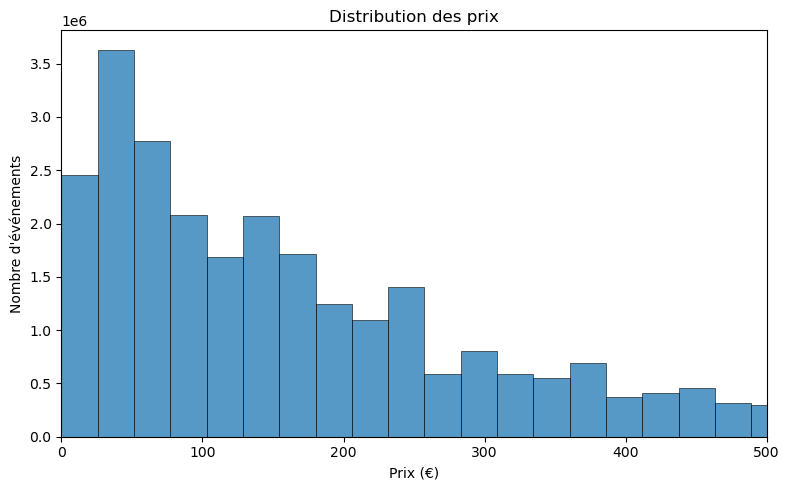


Figure 7 - Distribution des prix

* + 1. **Présence de valeurs nulles**

L’analyse des valeurs manquantes révèle que la plupart des colonnes sont complètes, à l’exception de category\_code et brand, qui présentent des absences régulières. Ces colonnes correspondent à des métadonnées de produits et peuvent ne pas être systématiquement renseignées selon les articles. Les autres champs, notamment event\_time, event\_type, user\_id et price, sont bien renseignés.

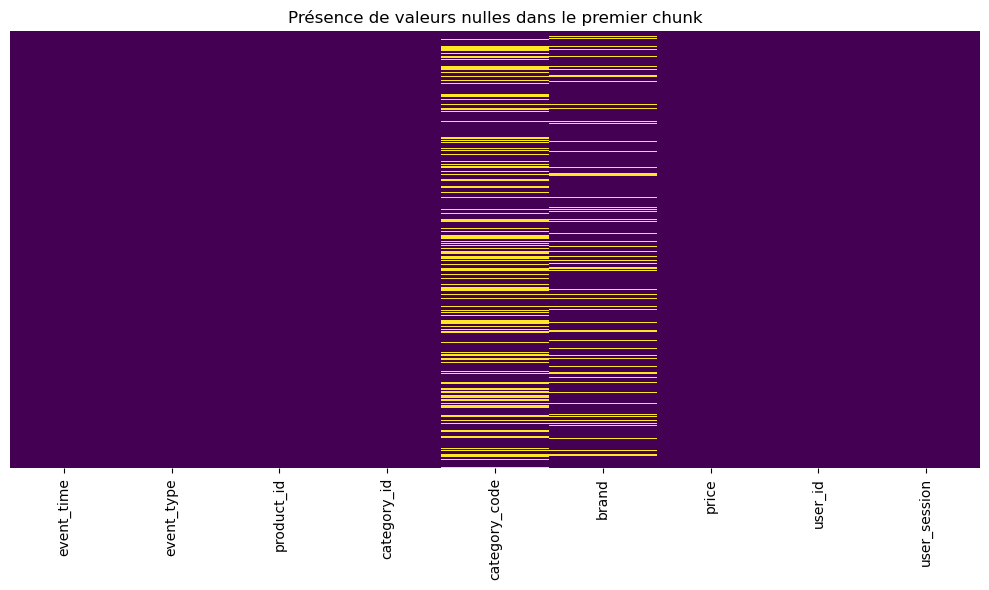


Figure 8 - Présence de valeurs nulles

1. **Stratégie de nettoyage**
   1. **Traitement des valeurs manquantes**

Les colonnes essentielles event\_time et user\_id, lorsqu’elles étaient nulles, ont été utilisées comme critère d’exclusion immédiate des lignes correspondantes. Les champs brand et category\_code, fréquemment incomplets mais utiles, ont été remplis par la valeur unknown pour conserver les données tout en signalant l’absence d’information. Pour la colonne price, les valeurs nulles ont été remplacées par 0 sauf dans le cas d’un achat (purchase) où l’absence de prix rendait l’opération invalide, entraînant donc leur suppression.

* 1. **Formatage et types de données**

Le champ event\_time a été converti au format datetime avec gestion des erreurs afin de garantir l’exploitabilité temporelle. Les champs brand, category\_code et user\_session ont été castés en str pour uniformiser leur traitement. Le champ price, potentiellement mal formaté, a été converti en numérique (float) avec traitement des erreurs, assurant la cohérence des calculs futurs.

* 1. **Nettoyage des valeurs aberrantes**

Les valeurs de prix inférieures ou égales à zéro, bien qu'autorisées pour des actions comme la visualisation ou l’ajout au panier, ont été conservées sous forme de 0 uniquement lorsqu’il ne s’agissait pas d’un achat. Les achats sans prix valide ont été supprimés, car ils ne peuvent être interprétés correctement dans une analyse comportementale liée aux dépenses.

* 1. **Filtrage des événements non pertinents**

Le champ event\_type a été filtré pour ne conserver que les événements utiles à l’analyse des comportements d’achat : view, cart, remove\_from\_cart et purchase. Les autres types d’événements, s’ils existent, ont été ignorés pour simplifier l’analyse et concentrer la modélisation sur les interactions réellement significatives.

* 1. **Séparation des catégories produits**

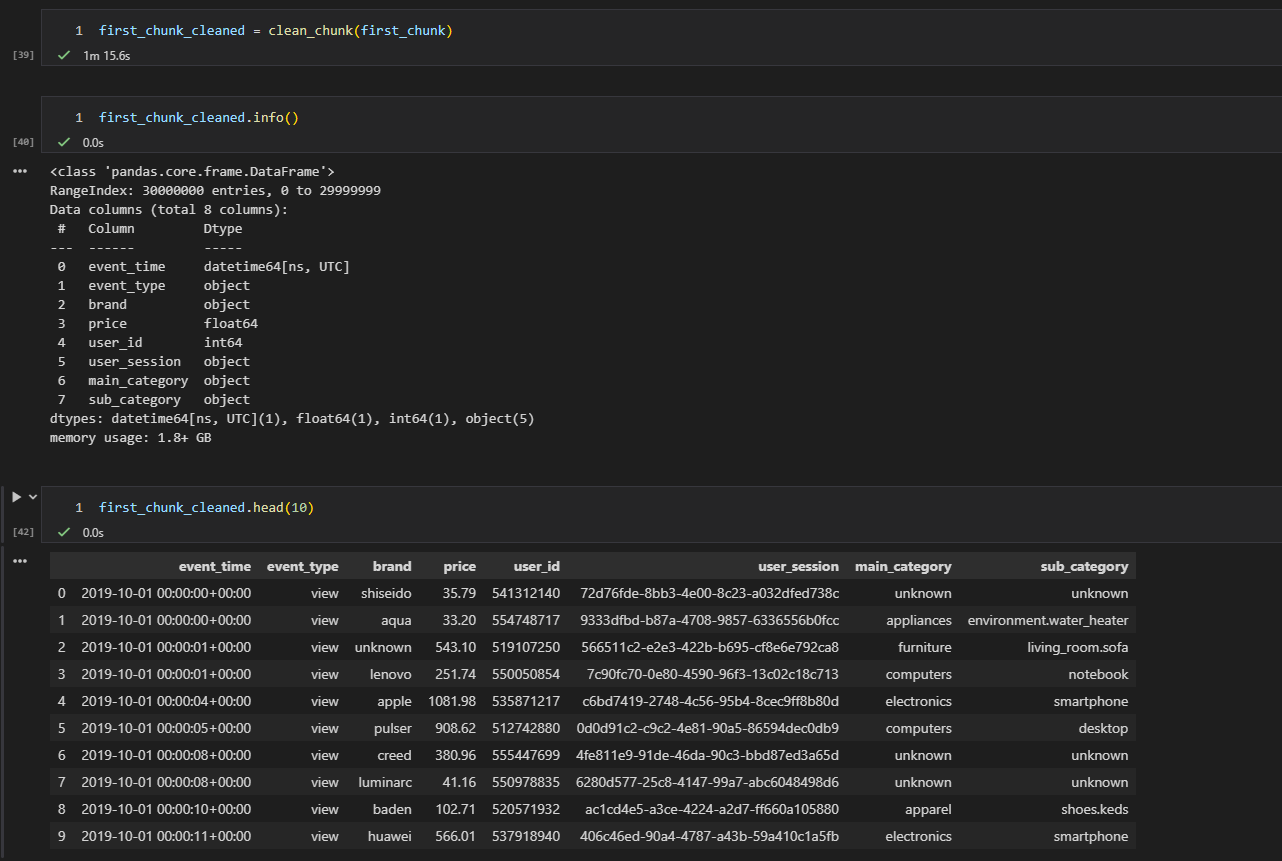
La colonne category\_code, composée de catégories hiérarchiques au format main.sub, a été scindée en deux colonnes distinctes : main\_category pour le niveau principal et sub\_category pour le détail. Ce découpage permet une meilleure lisibilité et une analyse plus fine des préférences produits, même lorsque certaines catégories sont absentes ou mal renseignées.

* 1. **Finalisation du jeu de données**

Enfin, les colonnes jugées inutiles pour la suite du traitement, à savoir category\_id, category\_code et product\_id, ont été supprimées. L’index a été réinitialisé pour garantir une structure propre, et le DataFrame nettoyé est prêt à être agrégé et analysé dans les étapes de feature engineering.

* 1. **Vérification après nettoyage**

Après application de la fonction de nettoyage, une inspection rapide a permis de valider l’intégrité du nouveau jeu de données. Les colonnes inutiles ont bien été supprimées, les types de données corrigés, et les champs sensibles comme event\_time, user\_id, event\_type ou price sont désormais propres et exploitables. Le tableau ne contient plus de valeurs nulles critiques, ce qui permet d’engager sereinement l’étape d’agrégation utilisateur. Ci-dessous les résultats.



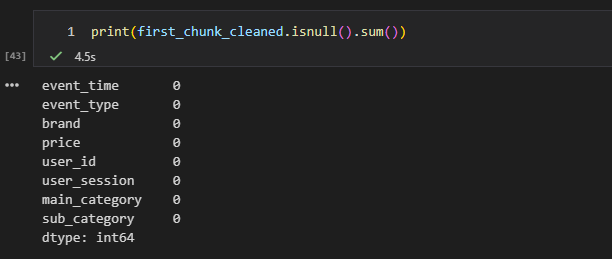


Figure 9 - Vérification des données nettoyées

* 1. **Sauvegarde des données nettoyées**

Après l'application du nettoyage sur chaque chunk de données brutes, les blocs traités ont été immédiatement sauvegardés dans un répertoire intermédiaire au format parquet. Ce format a été choisi pour sa capacité à stocker efficacement des données tabulaires, tout en réduisant significativement le volume occupé sur le disque. Il permet également une lecture et une réutilisation plus rapides que les fichiers CSV.

Chaque chunk nettoyé a été exporté dans un dossier interim. Cette démarche a permis d’éviter de relancer le nettoyage complet à chaque étape du projet, de conserver une trace des données intermédiaires fiables, et de faciliter le développement des étapes suivantes sur un volume réduit et maîtrisé.

Le volume total des fichiers nettoyés a ainsi été ramené à environ 13 Go, contre 52 Go pour les données initiales. Cette compression résulte à la fois de l’allègement des colonnes (suppression, typage, nettoyage) et de l’efficacité du format parquet.

1. **Construction du jeu utilisateur**
   1. **Objectif de l’agrégation**

Cette étape vise à synthétiser le comportement des utilisateurs à partir des événements qu’ils ont générés sur la plateforme Amazing. Plutôt que d’analyser les événements ligne par ligne, les données sont agrégées par user\_id afin de générer un profil comportemental représentatif pour chaque client.

* 1. **Approche progressive et sécurisée**

Un premier traitement a été effectué sur un seul fichier nettoyé, afin de valider la logique d’extraction des caractéristiques (features). Une fois le résultat jugé cohérent, le traitement a été généralisé à l’ensemble des fichiers parquet du dossier data/interim/, avec export dans data/processed/. Cette approche permet à la fois la reproductibilité, la traçabilité et la reprise facile en cas d’interruption.

* 1. **Caractéristiques extraites**

Pour chaque utilisateur, les métriques suivantes ont été calculées :

* Comptage des événements par type : vues, ajouts au panier, achats
* Nombre de sessions uniques
* Catégories préférées : principale et secondaire (calculées par fréquence)
* Durée d’activité : nombre de jours actifs
* Montants financiers : total dépensé et panier moyen
* Récence : nombre de jours depuis le dernier événement
* Périodes clés : date du premier et du dernier événement

Ces variables sont issues d’un agrégat groupby(user\_id) enrichi par des jointures avec un sous-ensemble purchase\_df pour les calculs financiers.

* 1. **Architecture de traitement**

Le traitement a été réalisé en mode batch, fichier par fichier, avec journalisation (logger) à chaque étape. Chaque bloc de résultats est sauvegardé dans un fichier user\_features\_chunk\_X.parquet dans data/processed/. Le log final confirme que 17 chunks ont été traités avec succès, pour un total d’environ 6h23 d’exécution.

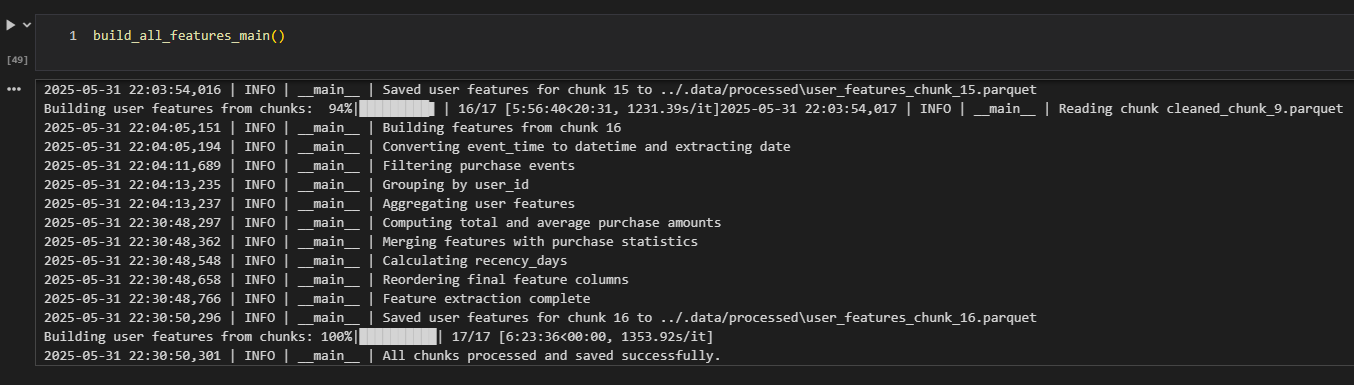


Figure 10 - Architecture de traitement des données en batch

* 1. **Fusion finale des utilisateurs multi-chunks**

Après l'extraction des caractéristiques utilisateur sur chacun des 17 fichiers traités individuellement, une fusion complète a été réalisée pour regrouper les données par user\_id. En effet, un même utilisateur pouvait apparaître dans plusieurs fichiers, notamment s’il avait été actif sur plusieurs périodes. Une étape d’agrégation finale a donc été appliquée pour consolider toutes ses interactions en une seule ligne.

Chaque variable a été agrégée selon une logique adaptée : les fréquences d’événements et le total dépensé ont été additionnés (sum), les dates de première et dernière activité ont été respectivement conservées via les minimums et maximums (min, max), la catégorie préférée a été déterminée à partir du mode statistique, et le panier moyen a été calculé via la moyenne des prix (mean). Cette étape permet d’assurer une vue complète et non redondante de chaque profil utilisateur.

Le résultat a été sauvegardé dans le fichier unique user\_features.parquet, situé dans le dossier data/processed/. Ce fichier regroupe **15 639 803 utilisateurs uniques** pour un volume final d’environ **390 Mo**, soit une **réduction de plus de 99 % par rapport aux 52 Go de données brutes initiales**. Ce format compact et structuré permet désormais une exploration fluide, même sur des machines à ressources limitées, tout en préservant la richesse des comportements clients nécessaires à la modélisation.

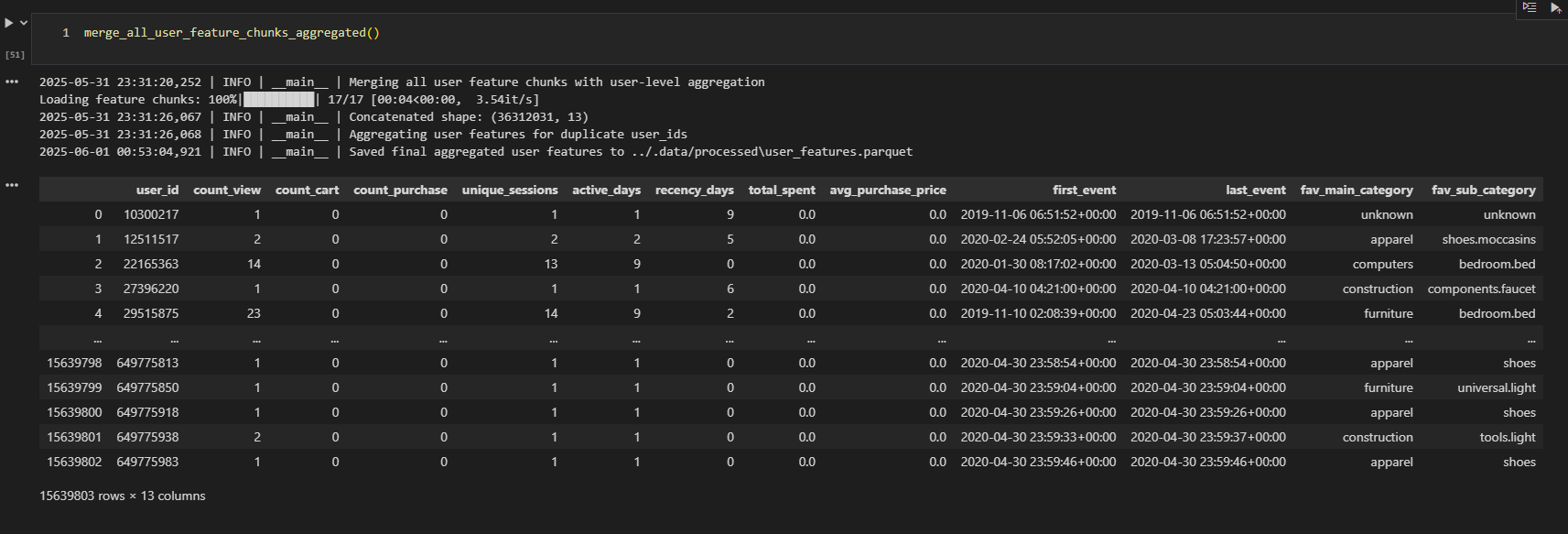


Figure 11 - Résultat de la fusion des données en un seul dataframe

1. **Analyse exploratoire**
   1. **Objectifs de l’exploration**

L’analyse exploratoire des données (EDA) constitue une étape essentielle dans tout projet de modélisation. Elle permet de comprendre la structure du jeu de données final, de détecter d’éventuelles incohérences, et d’identifier des patterns comportementaux pertinents. Dans le cadre de ce projet, l’objectif est de dégager des profils utilisateurs représentatifs à partir des variables agrégées, tout en évaluant la qualité et la distribution des données issues du nettoyage et de la transformation précédente.

* 1. **Statistiques générales des utilisateurs**

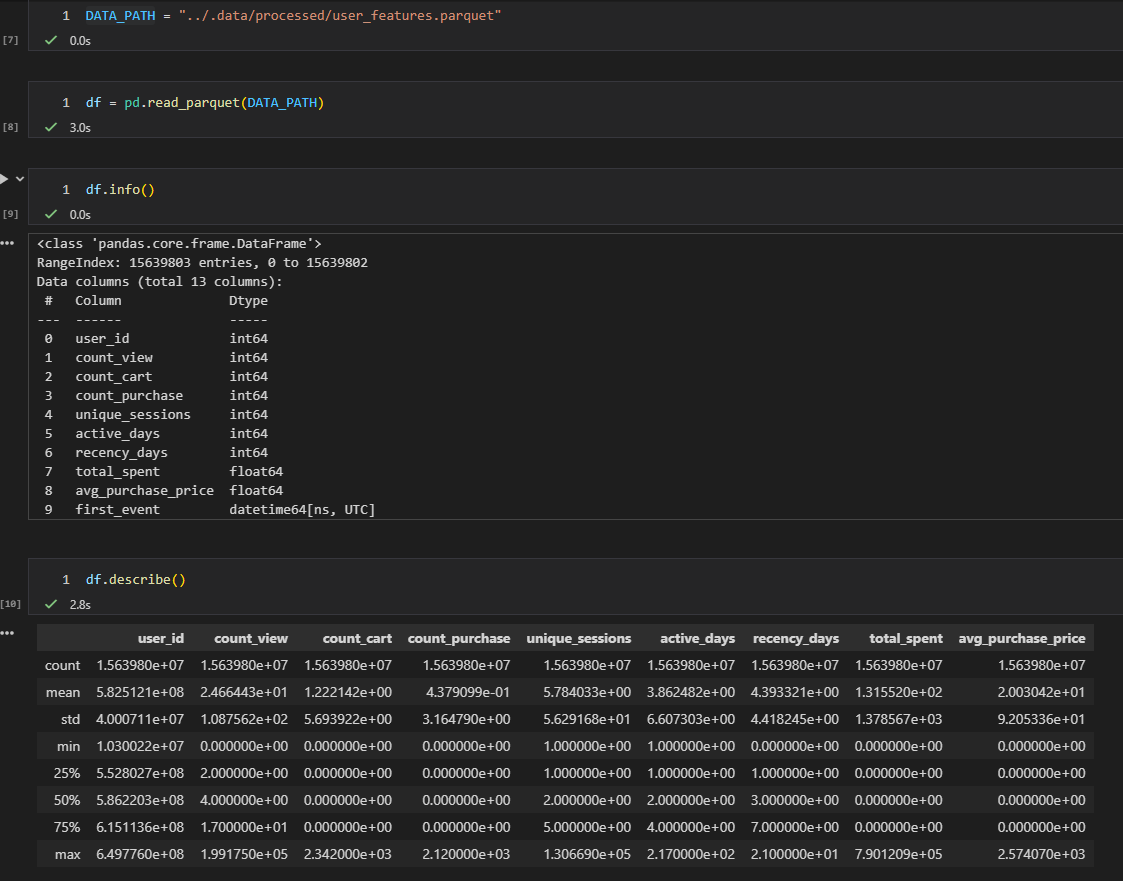


Figure 12 - Statistiques générales des utilisateurs

L’analyse statistique des données agrégées des utilisateurs révèle une forte hétérogénéité dans les comportements. En moyenne, un utilisateur a consulté environ 24 produits, ajouté 1,2 article au panier, et effectué moins d’un achat. Toutefois, les médianes beaucoup plus faibles indiquent que la majorité des utilisateurs ont eu une activité très limitée. En effet, 75 % des utilisateurs n’ont réalisé aucun achat. Les dépenses moyennes sont proches de zéro pour une grande partie d’entre eux, bien que certains profils aient dépensé jusqu’à 790 000 €, ce qui suggère la présence de clients très actifs. Le nombre moyen de jours actifs est inférieur à 4, et la récence moyenne d’activité est de 4,3 jours, ce qui pourrait indiquer une faible fidélisation ou une forte proportion de visiteurs occasionnels. Ces observations confirment un déséquilibre entre utilisateurs passifs et quelques clients très engagés.

* 1. **Analyse des comportements d’interaction**

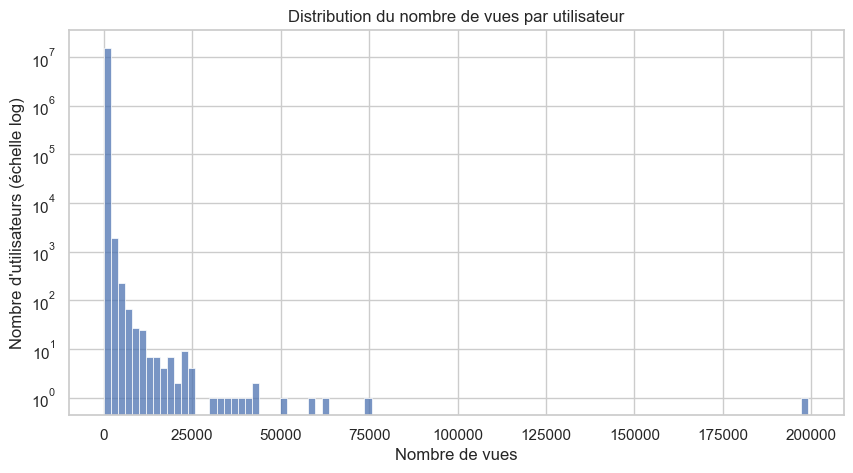


Figure 13 - Distribution du nombre de vues par utilisateur

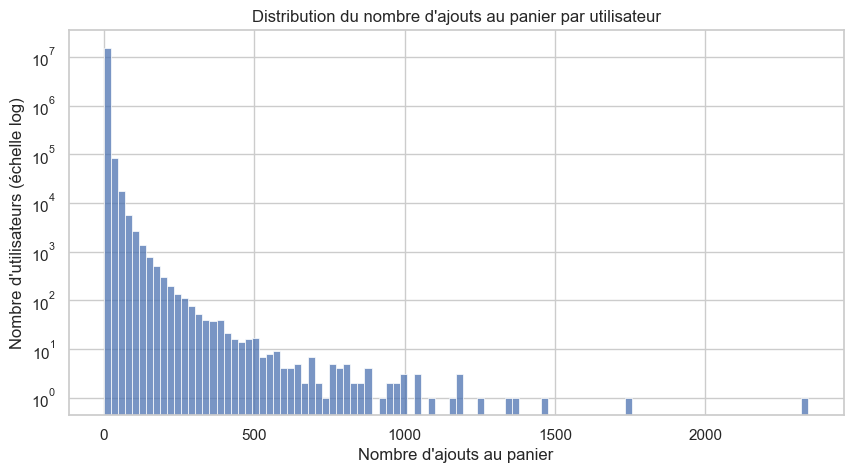


Figure 14 - Distribution du nombre d'ajouts au panier par utilisateur

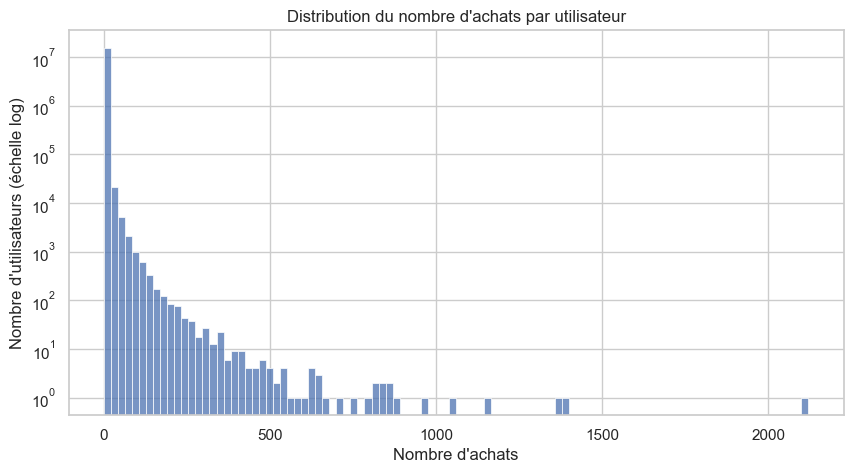


Figure 15 - Distribution du nombre d'achats par utilisateurs

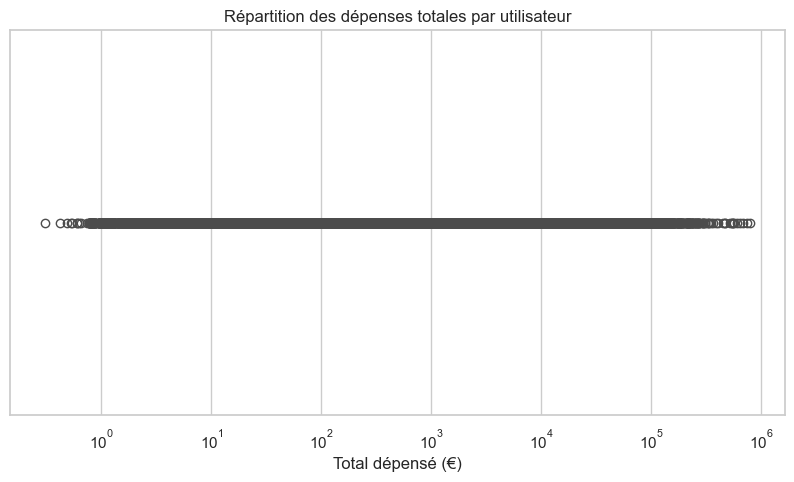


Figure 16 - Répartition des dépenses totales par utilisateur

Les distributions des variables count\_view, count\_cart et count\_purchase, présentées en échelle logarithmique, mettent en évidence une très forte concentration d’utilisateurs autour de valeurs faibles, combinée à une longue traîne de cas extrêmes. La majorité des utilisateurs ont consulté moins de 10 produits, n’ont ajouté aucun article au panier, et n’ont réalisé aucun achat. Ce phénomène, très visible sur les histogrammes, traduit un comportement globalement passif de la base client.

Cependant, certains profils atypiques ressortent : quelques utilisateurs ont dépassé les 100 000 vues ou ont effectué plus de 1000 achats. Ces cas rares, bien que marginaux numériquement, constituent des profils à haute valeur potentielle. Cette opposition entre passivité généralisée et hyperactivité minoritaire suggère qu’une segmentation pourrait utilement distinguer des groupes tels que : les simples visiteurs (beaucoup de vues, peu ou pas d’achats), les acheteurs sporadiques, les clients fidèles et récurrents.

* 1. **Variables d’engagement et récence**

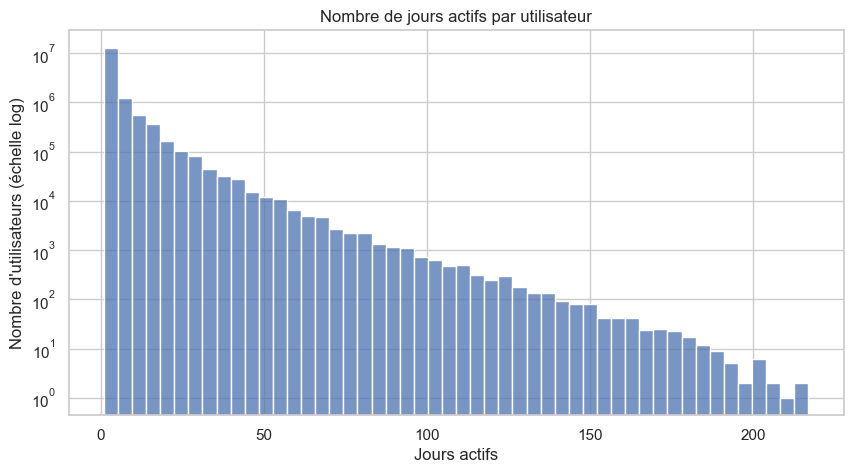


Figure 17 - Nombre de jours actifs par utilisateur

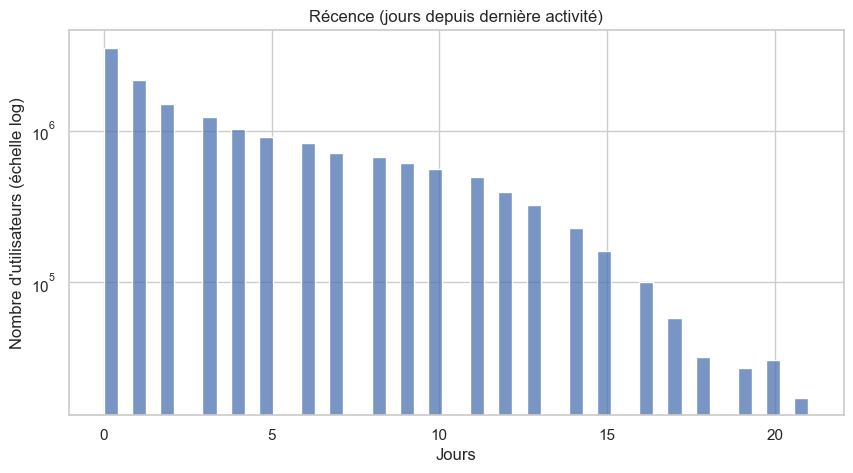


Figure 18 - Récence par utilisateur

Les variables active\_days, unique\_sessions (analysée en partie dans la structure), et recency\_days fournissent des informations précieuses sur la régularité et la fraîcheur des interactions.

L’histogramme du nombre de jours actifs révèle une structure décroissante : une majorité des utilisateurs ont été actifs une seule journée, et seul un faible pourcentage dépasse les 50 jours d’activité. Cette faible récurrence d’utilisation traduit une base client largement constituée d’utilisateurs ponctuels, voire de simples visiteurs uniques.

La variable recency\_days permet quant à elle d’évaluer la fraîcheur de l’engagement. Le graphique montre que la majorité des utilisateurs ont été actifs très récemment, souvent dans les 5 derniers jours. Ce pic récent pourrait s’expliquer par la période couverte dans les données, mais il met aussi en lumière l’existence d’un vivier d’utilisateurs à relancer ou à fidéliser. À l’inverse, les utilisateurs dont l’activité remonte à plus de 15 jours pourraient être considérés comme dormants.

* 1. **Préférences de produits**

L’analyse des catégories préférées par utilisateur permet d’identifier les pôles d’intérêt dominants sur la plateforme. La variable fav\_main\_category, calculée comme la catégorie principale la plus fréquente chez chaque utilisateur, montre une concentration marquée autour de quelques thématiques populaires.

Les dix catégories les plus fréquentes comprennent notamment construction, electronics, appliances, computers, et phones, ce qui suggère une forte polarisation des comportements vers des produits technologiques. D’autres catégories comme furniture, kids ou sport apparaissent plus marginalement.

Une segmentation future pourrait bénéficier de ces informations pour regrouper les utilisateurs selon leurs préférences thématiques, soit en intégrant ces variables encodées dans le modèle, soit en les croisant après clustering pour interpréter les profils.

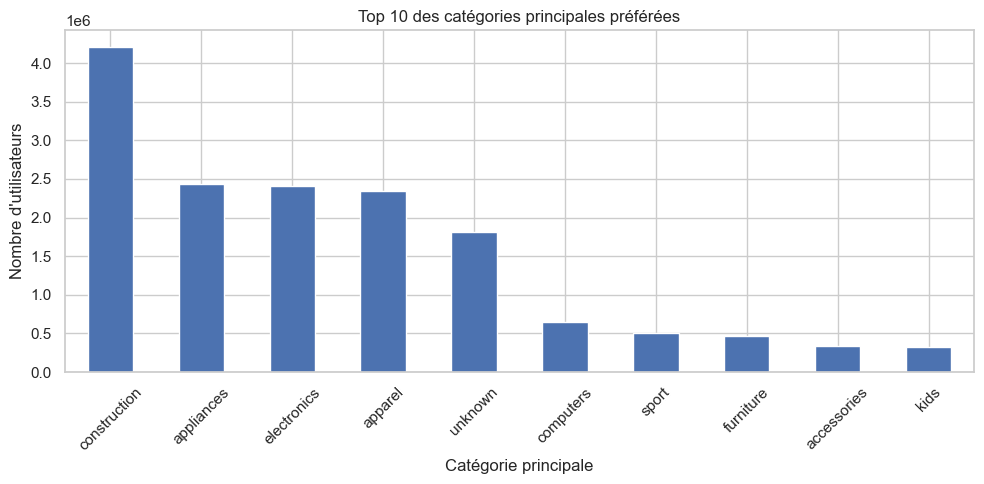
****

Figure 19 - Top 10 des catégories principales préférées

* 1. **Corrélations entre variables numériques**

Avant de procéder à une modélisation, il est essentiel d’évaluer les corrélations entre les variables numériques. Une matrice de corrélation permet d’identifier les liens éventuels entre les variables de fréquence (count\_view, count\_purchase, etc.), de valeur (total\_spent) et d’engagement (active\_days, recency\_days, etc.).

La heatmap obtenue montre que certaines variables sont naturellement corrélées, notamment count\_purchase et total\_spent, ce qui est attendu. D'autres variables, comme recency\_days ou fav\_main\_category, sont très peu corrélées aux autres, ce qui confirme leur valeur potentielle pour capturer une information complémentaire.

Cette analyse justifie la conservation de toutes les variables numériques dans le jeu de données à ce stade, en vue de la segmentation.

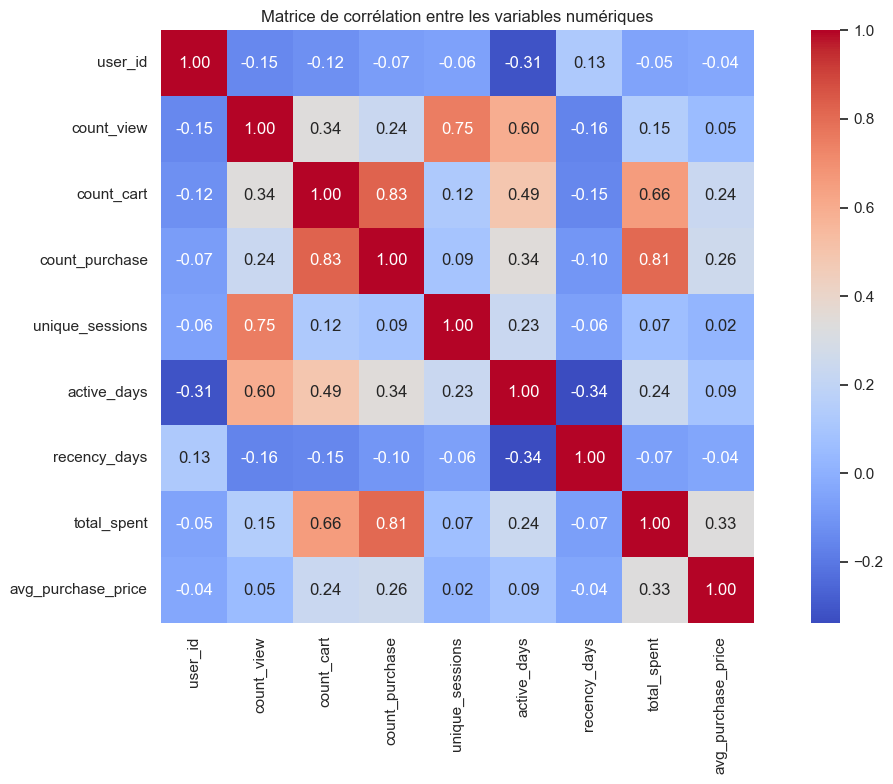


Figure 20 - Matrice de corrélation entre les variables numériques

**Conclusion**

Ce rapport a permis d’établir une base solide pour la compréhension et la structuration des données issues de la plateforme Amazing. À partir d’un volume initial de 52 Go répartis sur sept mois d’activité, un pipeline de traitement par blocs a été mis en place, comprenant le nettoyage, la transformation, l’agrégation comportementale par utilisateur, et la sauvegarde optimisée au format parquet.

L’exploration des données utilisateur a révélé des profils très hétérogènes : une majorité de comptes faiblement actifs et une minorité concentrant un volume élevé d’interactions et de dépenses. Les analyses statistiques et graphiques ont mis en lumière la forte asymétrie des comportements, les déséquilibres structurels dans l’usage de la plateforme, ainsi que la diversité des préférences produits.

Enfin, la qualité du jeu final, sa structure tabulaire par utilisateur, et la richesse des variables comportementales disponibles constituent désormais un point de départ optimal pour la mise en place d’une segmentation non supervisée. Celle-ci fera l’objet d’un second rapport exclusivement consacré à la modélisation.