

ÉVALUATION FINALE ATELIER SPARK SCALA

Réalisé par :

- Yann Boris KUETI
- Roseline MONTHE
- Carole FOBASSO
- Rebecca KOUAKOU

Présenté pour Mme : Amina MARIE

Date : 26 mai 2025

Table des matières

1	Introduction	2
2	Étape 1 : Prétraitement des données	2
3	Étape 2 : Analyse des données	2
4	Visualisation des Résultats avec XChart	5
4.1	Analyse des Visualisations	6
4.2	Recommandations et Perspectives Prédictives	7
5	Conclusion	8

1 Introduction

Dans le cadre de notre formation **d'expert en informatique et systèmes d'information** à l'EPPI, nous avons mené un projet complet de traitement et d'analyse de données e-commerce.

Ce travail a été réalisé en utilisant la puissance du **framework Apache Spark** combiné au langage **Scala**, afin de manipuler efficacement un grand volume de données issues d'une plateforme de vente en ligne. Ces données, brutes à l'origine, contiennent des informations précieuses sur les utilisateurs, les produits et les transactions.

L'objectif de ce projet est d'extraire des **informations pertinentes** permettant de mieux comprendre le comportement des clients, les performances des produits, ainsi que les habitudes d'achat. Pour ce faire, une démarche structurée a été adoptée, allant de la préparation des données à leur interprétation visuelle.

2 Étape 1 : Prétraitement des données

Cette étape visait à charger, nettoyer et préparer les données pour les étapes d'analyse ultérieures. Elle a été réalisée dans le fichier `Etape1TransformationData.scala`.

Objectif : Nettoyer et structurer les données pour permettre une analyse cohérente et efficace.

Actions réalisées :

- Chargement du fichier CSV original `ecommerce_data_enriched.csv` avec Spark.
- Filtrage des données incomplètes ou incohérentes.
- Filtrage des sessions inférieures à une minute.
- Création de nouvelles colonnes : heure, jour de la semaine, type de session (courte/longue selon `session_duration < 10`).
- Standardisation des valeurs : mise en minuscules des types d'appareils.
- Enregistrement des données nettoyées et transformées dans le dossier `data/nettoye.csv`.

Difficultés rencontrées et solutions :

- *Problème de format des dates/heures* : résolu grâce aux fonctions Spark `to_timestamp` et `to_date` pour assurer une conversion fiable.

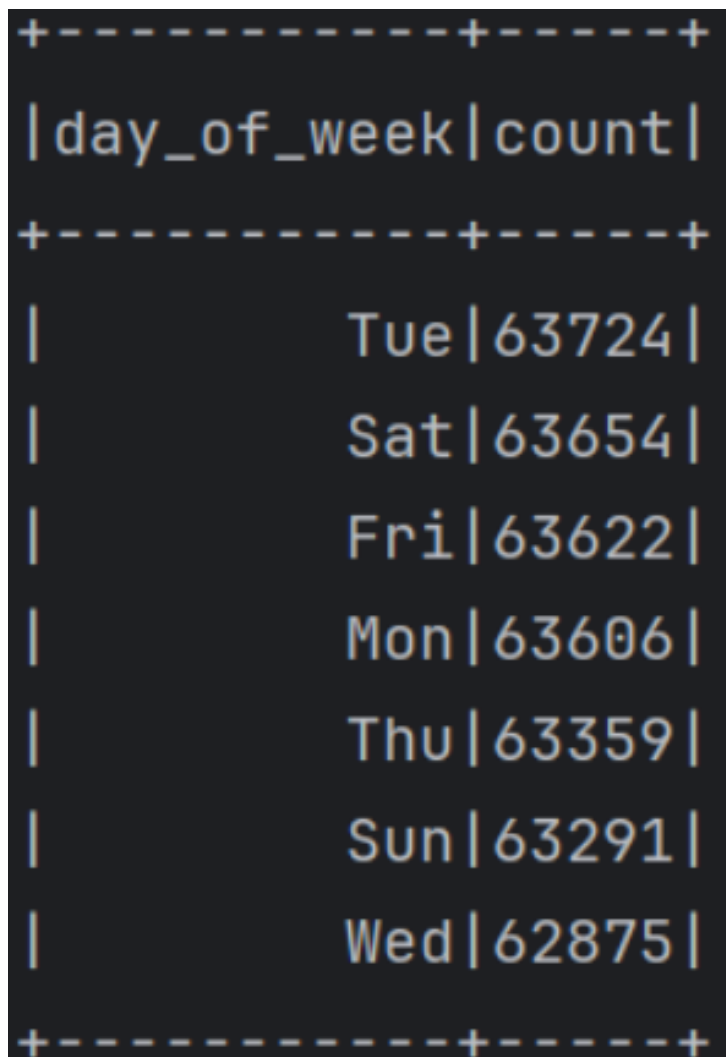
Aperçu du jeu de données nettoyé :

user_id	session_duration	pages_viewed	product_category	purchase_amount	review_score	review_text
U772383	8	1	High-tech	21.87	4.0	Excellent produit
U287378	5	6	Livres	39.53	3.0	Produit conforme
U471939	1	5	Beauté	7.23	1.0	Livraison rapide
U326713	2	8	Maison	149.89	3.0	Bon rapport quali...
U154646	11	3	Mode	36.56	2.0	Produit conforme
U916567	8	5	Maison	42.24	4.0	Parfait pour mes ...
U855496	11	7	High-tech	25.68	2.0	A éviter
U767477	5	4	Beauté	133.01	5.0	Livraison rapide
U404156	1	8	Mode	13.35	2.0	Pas comme décrit
U365067	23	9	Beauté	7.15	2.0	Bon achat

3 Étape 2 : Analyse des données

Nous avons effectué une série d'analyses statistiques pour comprendre les comportements des utilisateurs.

Distribution hebdomadaire des sessions

A terminal window with a dark background and light-colored text. It displays a table with two columns: 'day_of_week' and 'count'. The table is enclosed in a dashed border. The data rows are: Tue (63724), Sat (63654), Fri (63622), Mon (63606), Thu (63359), Sun (63291), and Wed (62875).

day_of_week	count
Tue	63724
Sat	63654
Fri	63622
Mon	63606
Thu	63359
Sun	63291
Wed	62875

Cette analyse montre que le mardi est le jour où il y a eu le plus de sessions enregistrées.

Répartition des notes clients

```
+-----+-----+
|review_score|count|
+-----+-----+
|          1.0|88685|
|          2.0|89039|
|          3.0|88409|
|          4.0|88794|
|          5.0|89204|
+-----+-----+
```

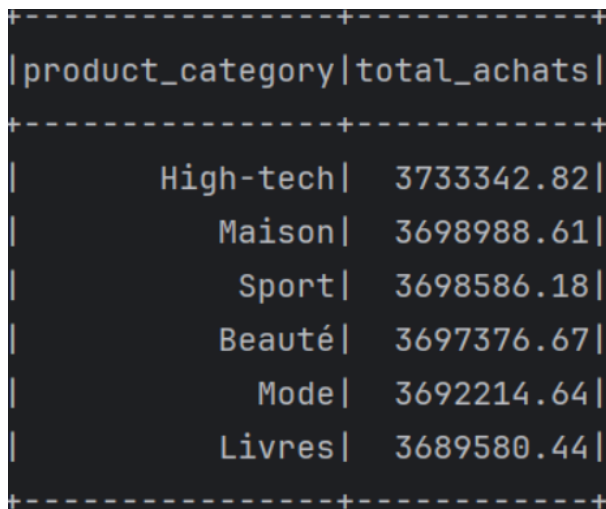
Les scores sont bien répartis avec une légère dominance de la note maximale (5.0).

Moyenne des notes par pays

```
+-----+-----+
|country|note_moyenne|
+-----+-----+
|Canada|3.01|
|France|3.0|
|Suisse|3.0|
|Tunisie|3.0|
|Maroc|3.0|
|Belgique|3.0|
+-----+-----+
```

La moyenne des évaluations reste stable à 3.0 pour la majorité des pays.

Chiffre d'affaires par catégorie de produit



```
+-----+-----+
|product_category|total_achats|
+-----+-----+
|      High-tech| 3733342.82|
|        Maison| 3698988.61|
|         Sport| 3698586.18|
|        Beauté| 3697376.67|
|         Mode| 3692214.64|
|        Livres| 3689580.44|
+-----+-----+
```

A terminal window displaying a table with two columns: 'product_category' and 'total_achats'. The table lists six product categories and their corresponding total purchase amounts. The 'High-tech' category has the highest value at 3,733,342.82.

product_category	total_achats
High-tech	3733342.82
Maison	3698988.61
Sport	3698586.18
Beauté	3697376.67
Mode	3692214.64
Livres	3689580.44

La catégorie High-tech est la plus lucrative du site.

Répartition des types d'appareils utilisés



```
+-----+-----+
|device_type|count|
+-----+-----+
|    mobile|310687|
|   desktop|133444|
+-----+-----+
```

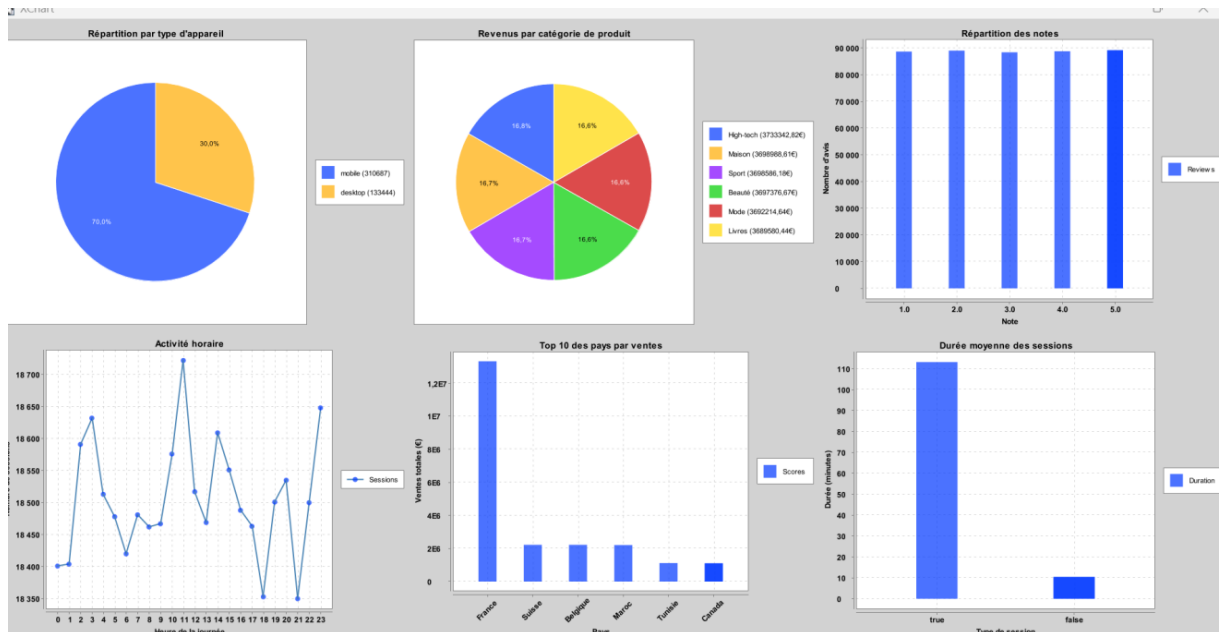
A terminal window displaying a table with two columns: 'device_type' and 'count'. The table shows the distribution of device types used by users. 'mobile' has a count of 310,687, and 'desktop' has a count of 133,444.

device_type	count
mobile	310687
desktop	133444

Le mobile représente plus de 70% des appareils utilisés.

4 Visualisation des Résultats avec XChart

Nous avons importé le fichier nettoyé dans XChart afin de générer des visualisations interactives et dynamiques. XChart a permis une exploration fluide des dimensions clients, produits et temporelles. Les graphiques suivants mettent en évidence des tendances clés et facilitent l'interprétation des comportements d'achat.



4.1 Analyse des Visualisations

1. Répartition par type d'appareil

lem

- *Observation* : 70% des sessions proviennent de mobiles contre 30% depuis desktop.
- *Interprétation* : L'expérience mobile est critique, suggérant d'optimiser le parcours d'achat sur smartphone (UX, performance).

2. Revenus par catégorie de produit

lem

- *Observation* : Les catégories se partagent équitablement 16–17% chacune, avec une légère dominance pour la High tech.
- *Interprétation* : Diversification des ventes équilibrée. Le marketing devrait renforcer les catégories moins dynamiques (ex. Livres, Mode) via promotions ciblées.

3. Répartition des notes clients

lem

- *Observation* : Distribution quasi normale, centrée autour de 4–5 étoiles.
- *Interprétation* : Globalement satisfaction élevée. Traquer toutefois les 1–2 étoiles pour identifier points de friction et améliorer le service.

4. Activité horaire

lem

- *Observation* : Pic de sessions entre 11h et 13h, suivi d'une légère baisse, puis un rebond en fin de journée (17h–19h).
- *Interprétation* : Planifier les campagnes e-mails ou promotions flash aux heures de pointe, notamment vers midi et début de soirée.

5. Top 10 des pays par ventes

lem

-
- *Observation* : La France domine largement (13 M), suivie de loin par la Suisse et la Belgique (2 M).
 - *Interprétation* : Renforcer la logistique et le support en France. Étudier des partenariats ou campagnes locales pour les marchés secondaires afin d'augmenter leur part.

6. Durée moyenne des sessions lem

- *Observation* : Les sessions longues (> 60 min) sont beaucoup moins fréquentes que les courtes.
- *Interprétation* : Le tunnel d'achat est rapide ; opportunité d'engager davantage l'utilisateur (contenu, recommandations, upselling) durant les sessions courtes.

4.2 Recommandations et Perspectives Prédictives

- **Optimisation Mobile** : Déployer A/B tests sur l'interface mobile pour réduire le taux d'abandon. *Prédiction* : une amélioration de la performance mobile de 20% pourrait augmenter les ventes mobiles de 10–15%.
- **Promotions Ciblées** : Lancer des offres flash à midi et en début de soirée. *Prédiction* : ces interventions, alignées sur les pics d'activité, pourraient accroître les conversions de 5% sur ces créneaux.
- **Segmentation Géographique** : Utiliser un modèle de scoring pour identifier les segments à fort potentiel (Belgique, Suisse). *Recommandation* : investir dans la publicité géo-ciblée pour augmenter ces marchés de 20% en 6 mois.

Problème de Gestion des Dates/Heures lors de la Visualisation

Lors de la création des graphiques temporels (notamment pour le graphique d'activité horaire), nous devons agréger et visualiser correctement les données datées extraites lors de l'étape de prétraitement (nettoyage et transformation).

Solution dans Etape3Visualizations.scala :

```
def createHourlyActivityChart(data: Array[(Int, Long)]): XYChart = {  
  val chart = new XYChartBuilder()  
    .width(900).height(500)  
    .title("Activité horaire")  
    .xAxisTitle("Heure de la journée")  
    .yAxisTitle("Nombre de sessions")  
    .build()  
  
  // Tri explicite des heures (0-23)  
  val sortedData = data.sortBy(_._1)  
  
  val series = chart.addSeries("Sessions",  
    sortedData.map(_._1.toDouble),  
    sortedData.map(_._2.toDouble))  
  
  // Configuration visuelle  
  series.setMarker(SeriesMarkers.CIRCLE)
```

```
series.setLineColor(new Color(70, 130, 180)) // Bleu acier
chart.getStyler.setXAxisMin(0.0)
chart.getStyler.setXAxisMax(23.0)
chart.getStyler.setXAxisTickMarkSpacingHint(1)

chart
}
```

Points Techniques Importants :

- **Tri des Heures** : `data.sortBy(_. _1)` garantit l'ordre chronologique (0h → 23h) et évite les graphiques avec les heures dans le désordre.
- **Formatage de l'Axe X** : `setXAxisMin/Max` force l'affichage de toutes les heures, empêchant la troncature automatique.
- **Marqueurs Temporels** : `setXAxisTickMarkSpacingHint(1)` affiche un libellé pour chaque heure.
- **Couleur et Style** : Ligne bleue avec marqueurs ronds pour une meilleure lisibilité.

5 Conclusion

Ce projet nous a permis de consolider nos compétences en Scala et Spark, tout en appliquant une méthodologie rigoureuse d'analyse de données. Nous avons identifié les grandes tendances commerciales et comportementales des utilisateurs, puis valorisé ces informations à travers des graphiques clairs. Pour améliorer ce travail, nous pourrions aller plus loin en introduisant des modèles prédictifs ou une segmentation client.