Université Claude Bernard Lyon 1

Behavioral Segmentation and Anomaly analysis

Khalil Ferhati

2025

Résumé

Ce projet a pour objectif d'analyser un ensemble de transactions bancaires afin d'identifier des comportements potentiellement frauduleux. Pour ce faire, des techniques de clustering et de détection d'anomalies ont été appliquées sur un jeu de données représentatif. L'analyse repose sur une approche d'Exploration de Données (EDA) approfondie permettant d'extraire des tendances et de comprendre la structure du dataset.

Table des matières

1 Introduction

L'industrie musicale a connu des transformations majeures avec l'avènement des plateformes de streaming comme Spotify, qui offrent un accès sans précédent à une vaste bibliothèque musicale et aux métadonnées associées. Comprendre les tendances musicales actuelles, les caractéristiques qui rendent un morceau populaire, et les dynamiques de collaborations entre artistes est essentiel pour les acteurs du secteur. Ce rapport vise à analyser les données musicales fournies par l'API Spotify afin de mettre en lumière ces aspects. Nous aborderons la présentation des données, leur nettoyage et préparation, une analyse exploratoire des tendances musicales, une segmentation des morceaux via le clustering, et une analyse des communautés artistiques à travers les graphes.

2 Présentation des Données

2.1 Description des variables

Voici les principales variables présentes dans le jeu de données :

- ***TransactionID* :** identifiant unique de chaque transaction.
- **AccountID :** identifiant du compte associé à la transaction.
- **TransactionAmount :** montant de la transaction.
- **TransactionDate :** date et heure de la transaction.
- **TransactionType :** type de transaction (Credit ou Debit).
- **Location :** lieu géographique de la transaction.
- **DeviceID :** identifiant de l'appareil utilisé.
- **IP Address :** adresse IP liée à la transaction.
- **MerchantID :** identifiant du commerçant.
- **Channel :** canal utilisé (en ligne, guichet, distributeur, etc.).
- **AccountBalance :** solde du compte après la transaction.
- **TransactionDuration :** durée de la transaction (en secondes).
- **LoginAttempts :** nombre de tentatives de connexion avant la transaction.
- **CustomerAge :** âge du client.
- **CustomerOccupation :** profession du client.
- **PreviousTransactionDate :** date de la transaction précédente.

2.2 Caractéristiques générales

- **Nombre total de lignes :** 2 512
- **Nombre de colonnes :** 16
- **Types de données :** mélange de variables numériques, catégorielles et temporelles
- **Valeurs manquantes :** aucune valeur manquante

Le jeu de données est donc complet et prêt à être utilisé pour des analyses statistiques et de détection d'anomalies.

2.3 Résumé statistique

TransactionAmount

- Moyenne : **297.59**
- Écart-type : **291.95**
- Minimum : **0.27**
- Maximum : **1919.11**
- **Observation :** La plupart des montants sont faibles, avec quelques transactions de valeur élevée (distribution asymétrique à droite).

CustomerAge

- Moyenne : **44.67 ans**
- Minimum : **18 ans**, Maximum : **80 ans**
- **Observation :** Le jeu de données couvre une large plage d'âges, utile pour segmenter les clients selon leur profil.

TransactionDuration

- Moyenne : **119.64 secondes**
- Minimum : **10 s**, Maximum : **300 s**
- **Observation :** Les durées varient selon le type de transaction. Les durées longues peuvent être liées à des transactions complexes ou suspectes.

LoginAttempts

- Moyenne : **1.12 tentatives**
- Maximum : **5 tentatives**
- **Observation :** La plupart des transactions ont une seule tentative de connexion, mais plusieurs tentatives peuvent indiquer une activité inhabituelle.

AccountBalance

- Moyenne : **5114.30 $**
- Minimum : **101.25 $**, Maximum : **14977.99 $**
- **Observation :** Les soldes varient fortement d'un compte à l'autre, ce qui reflète une clientèle diverse.

2.4 Observations principales

- **Accounts (comptes clients)**
  - 495 comptes différents sont présents dans les 2 512 transactions.
  - Certains comptes ont plusieurs transactions, comme le compte `AC00362` avec 12 transactions.

- **TransactionType**

- Les transactions **Debit** sont majoritaires, ce qui crée un déséquilibre entre les classes (Debit vs Credit).

- **Location (localisation)**

  - 43 lieux différents sont enregistrés.
  - La ville **Fort Worth** est la plus représentée avec 70 transactions.

- **DeviceID et IP Address**

  - 681 appareils et 592 adresses IP uniques ont été détectés.
  - Certains appareils et IP sont utilisés plusieurs fois, ce qui peut être un signe d'activité suspecte.

- **MerchantID (commerçant)**

  - 100 commerçants différents apparaissent dans les données.
  - Le commerçant `M026` a le plus grand nombre de transactions (45).

- **Channel (canal)**

  - Les transactions se font via trois canaux principaux : **en ligne, guichet et distributeur**.
  - Cette diversité aide à comparer les comportements selon le canal.

- **PreviousTransactionDate**

  - La présence de cette variable permet d'étudier le délai entre deux transactions d'un même client.

2.5 Interprétation et utilité

- Les variables comme **TransactionAmount**, **TransactionDuration**, **LoginAttempts** et **AccountBalance** sont particulièrement utiles pour la **détection d'anomalies**.

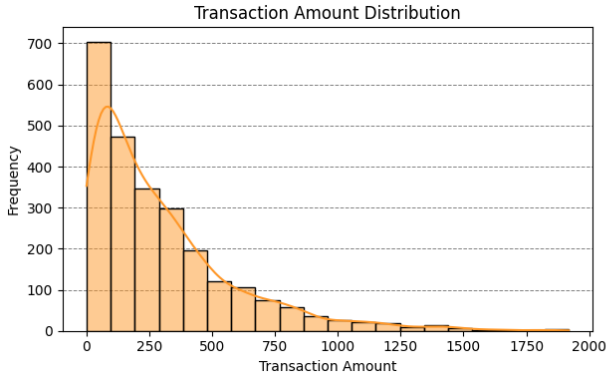
- Les variables **CustomerAge**, **CustomerOccupation** et **AccountBalance** permettent une **segmentation comportementale** des clients.

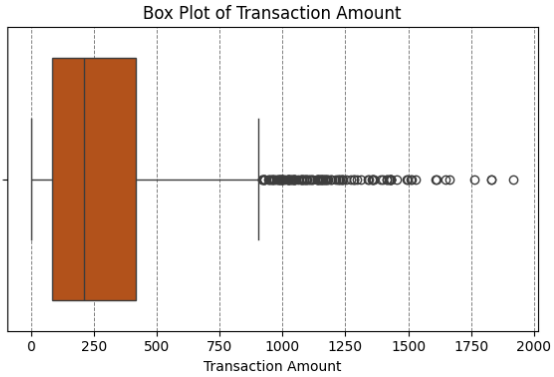
- L'analyse des dates et des durées aide à comprendre les **habitudes temporelles** et à repérer des changements de comportement.

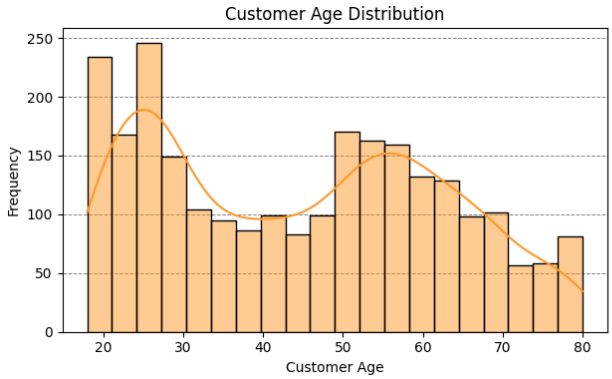
3 Analyse exploratoire des données (EDA)

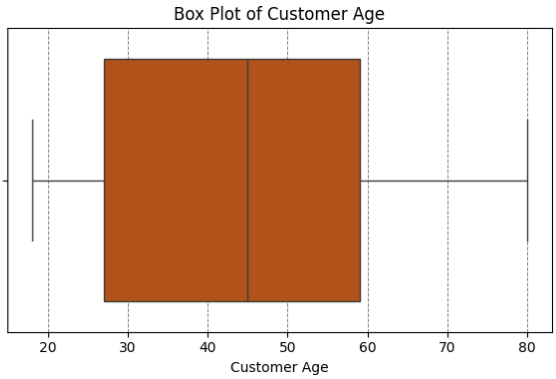
L'objectif de cette section est de mieux comprendre les tendances musicales, les caractéristiques des morceaux populaires et l'évolution des collaborations entre artistes. Cette analyse est appuyée par des visualisations pour illustrer les résultats clés.

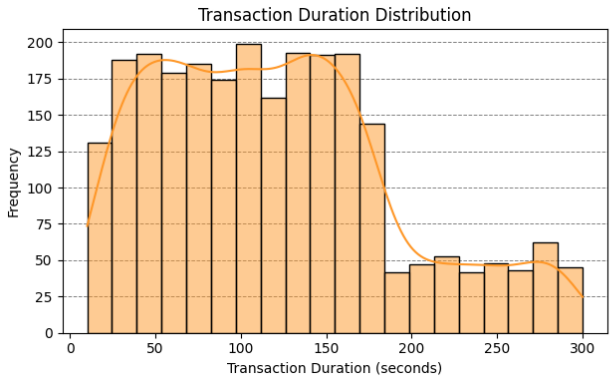
3.1 Analyse de distribution des colonnes

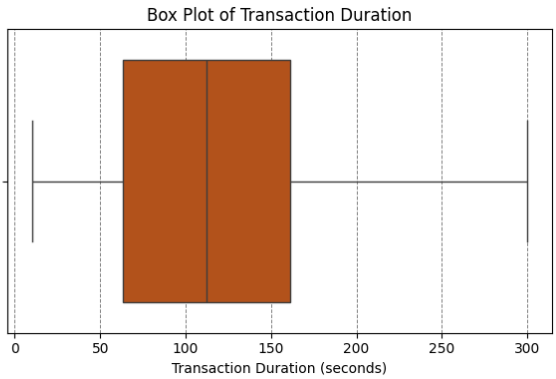


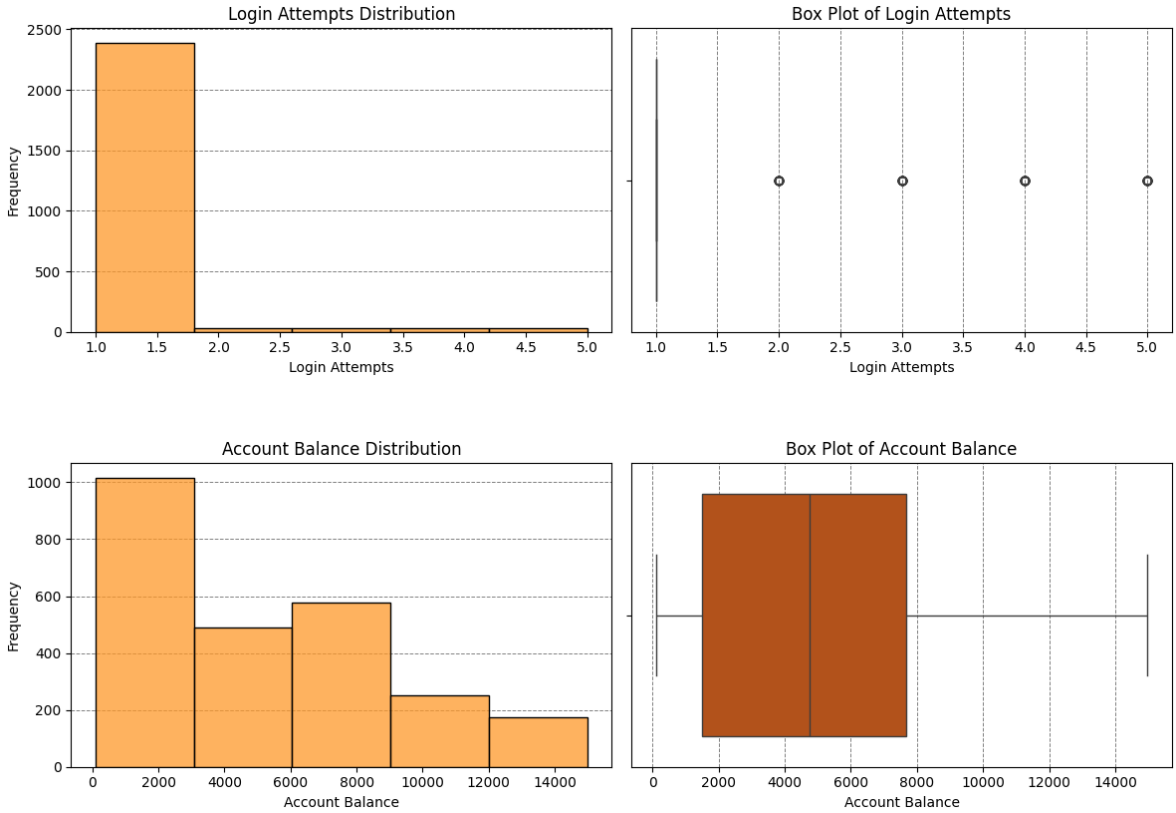










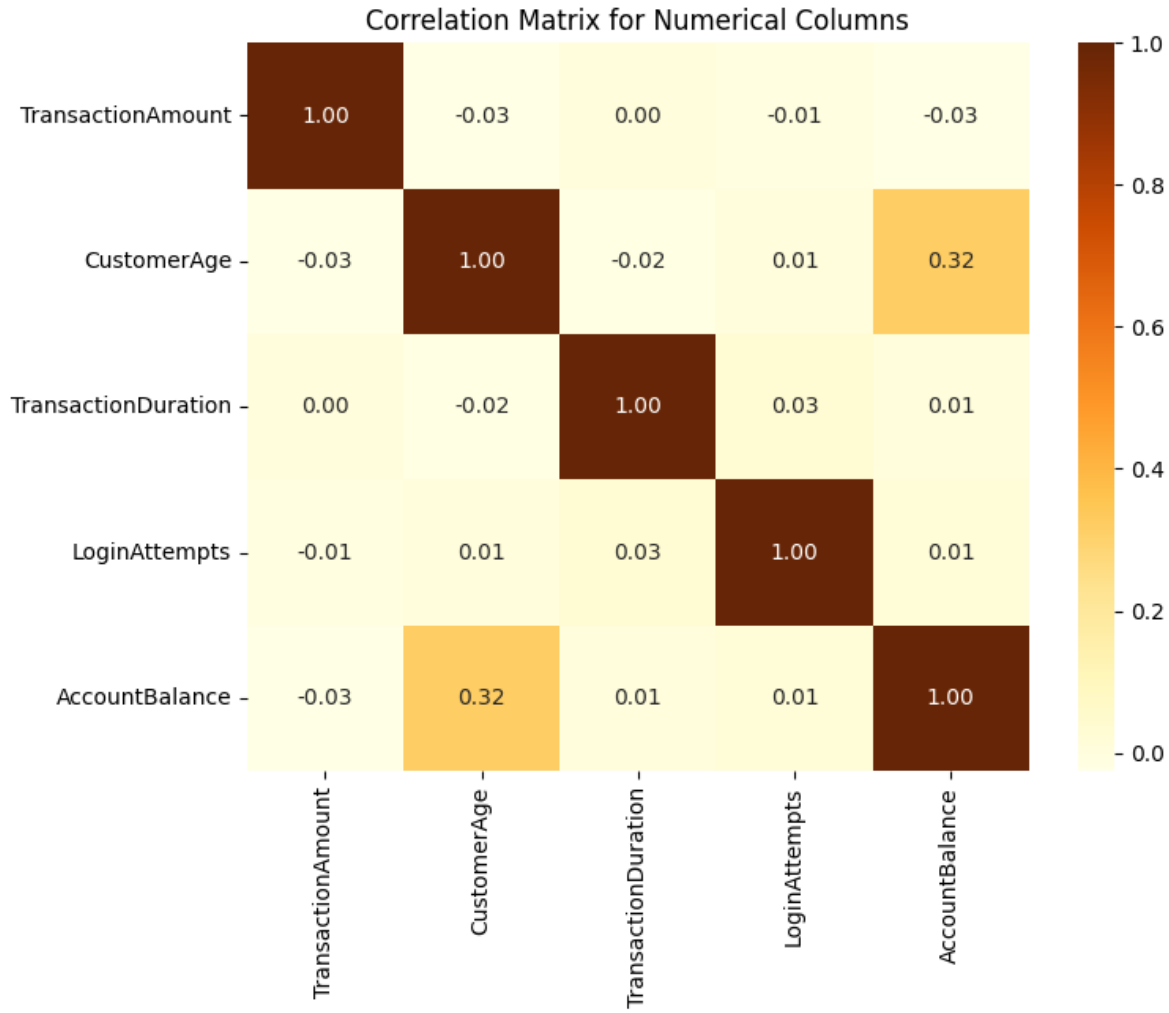


3.2 Analyse de la matrice de corrélation

1. **TransactionAmount et AccountBalance (corrélation positive)**
   On observe une corrélation positive modérée entre le montant de la transaction (`TransactionAmount`) et le solde du compte (`AccountBalance`).
   Cela signifie que les clients ayant un solde plus élevé ont tendance à effectuer des transactions plus importantes, ce qui correspond à un comportement financier attendu.
2. **CustomerAge (faible corrélation)**
   L'âge du client (`CustomerAge`) présente une corrélation faible, voire nulle, avec les autres variables numériques (`TransactionAmount`, `TransactionDuration`, `LoginAttempts`, `AccountBalance`).
   Cela indique que l'âge n'a pas d'influence directe sur le montant des transactions ou sur la durée des opérations.
3. **TransactionDuration et TransactionAmount (corrélation faible à modérée)**
   Une légère corrélation positive existe entre la durée d'une transaction et son montant.
   Les transactions plus longues peuvent être associées à des montants plus élevés, ce qui peut s'expliquer par des vérifications supplémentaires ou des processus de validation plus complexes.
4. **LoginAttempts (faible corrélation)**
   Le nombre de tentatives de connexion avant une transaction ne présente pas de corrélation notable avec les autres variables numériques.
   Cela suggère que ce paramètre est davantage lié au comportement individuel de l'utilisateur qu'à des tendances générales.
5. **AccountBalance et TransactionDuration (corrélation faible positive)**
   Une corrélation légèrement positive existe entre le solde du compte et la durée de la transaction.

Les comptes avec un solde plus élevé peuvent parfois impliquer des transactions plus longues, possiblement à cause de vérifications supplémentaires.

Observation générale :
La majorité des variables numériques présentent des corrélations faibles à modérées, indiquant une relative indépendance linéaire entre elles.

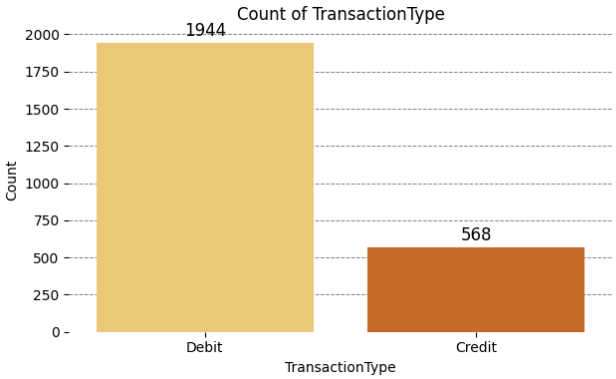


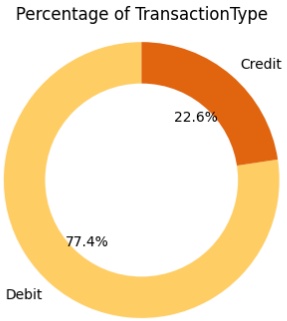
3.3 Analyse des transactions par type

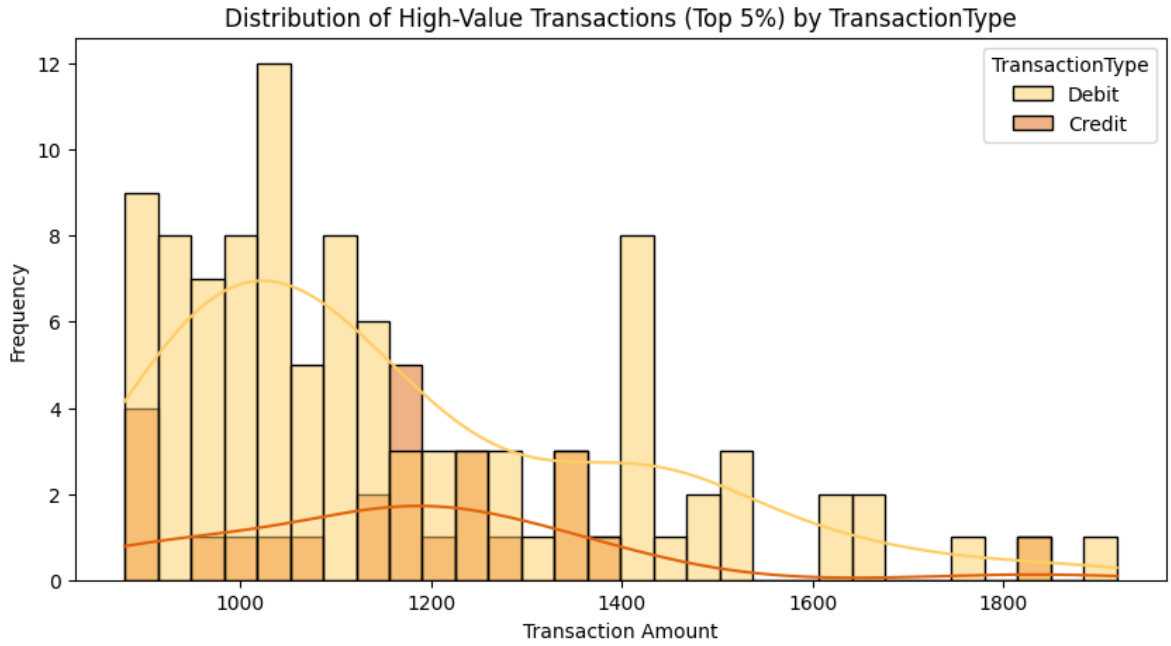
Le graphique des transactions montre que :

- Les **transactions "Debit"** sont largement majoritaires avec environ **1 944 opérations**.
- Les **transactions "Credit"** sont au nombre de **568**, représentant seulement **22,6 %** du total.

Les transactions "Debit" représentent environ **77,4 %** du jeu de données.
Ce déséquilibre entre les classes devra être pris en compte lors de la modélisation.





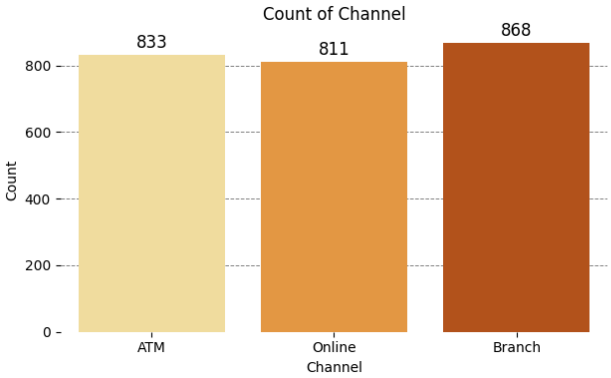


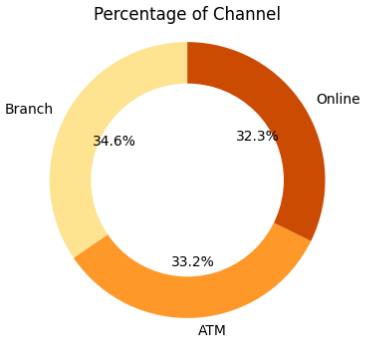
3.4 Répartition des transactions par canal

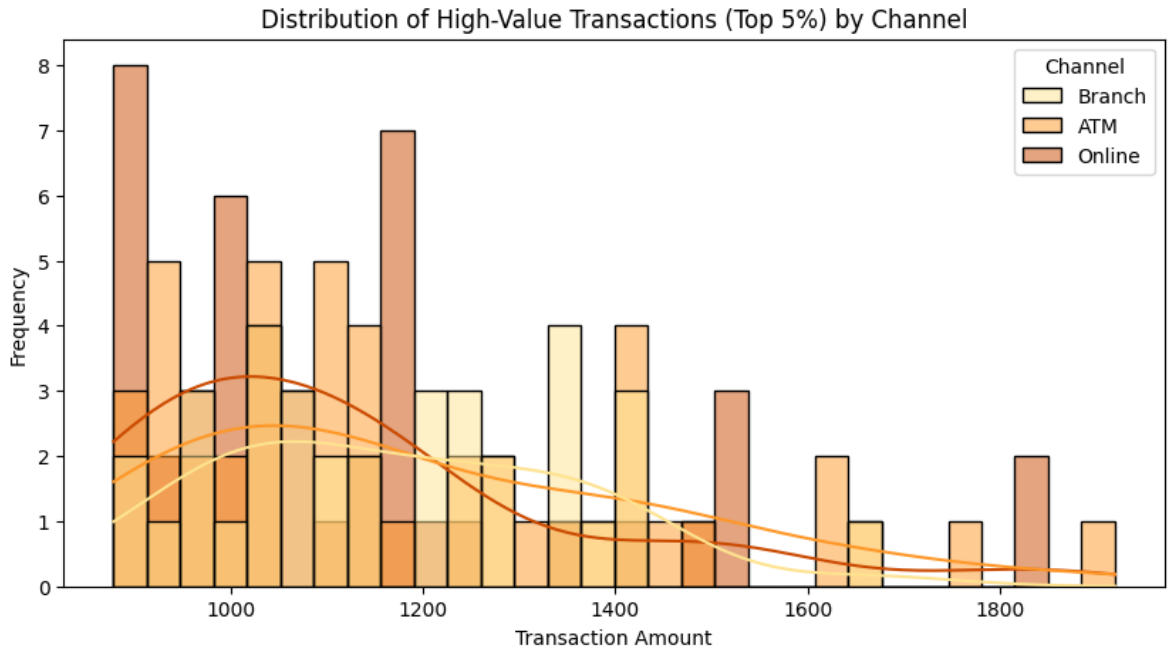
Les transactions sont réparties sur trois canaux : **ATM**, **Online**, et **Branch**.
Les graphiques montrent une distribution relativement équilibrée :

- Les transactions **en agence (Branch)** sont légèrement plus fréquentes.
- Les transactions via **ATM** et **Online** suivent de près, sans différences significatives.

Cela montre une utilisation variée des canaux par les clients.





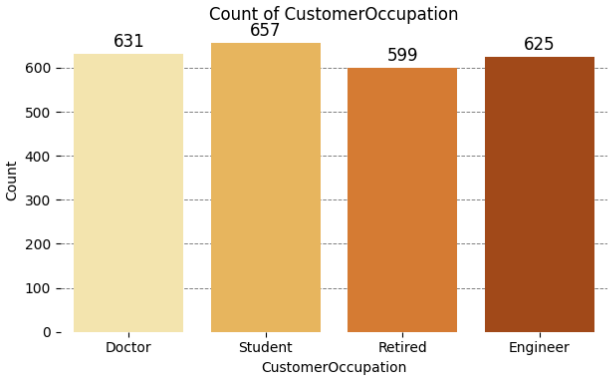


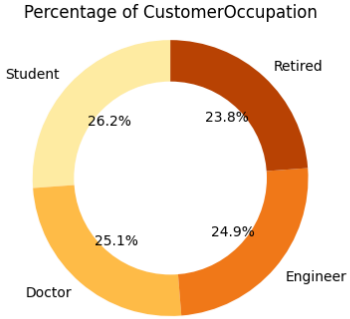
3.5 Répartition des transactions selon la profession

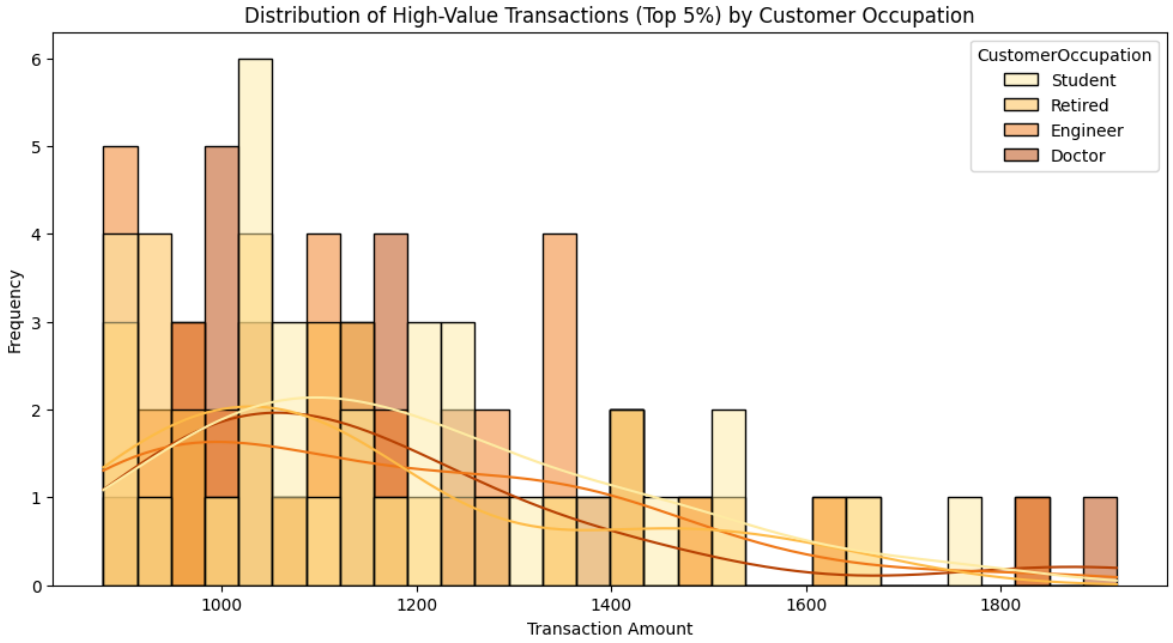
Le jeu de données contient quatre catégories de professions : **Doctor**, **Student**, **Retired**, et **Engineer**.
Les graphiques indiquent que :

- Les **étudiants (Student)** et les **médecins (Doctor)** effectuent le plus grand nombre de transactions.
- Les **retraités (Retired)** et **ingénieurs (Engineer)** présentent des volumes légèrement inférieurs.

Cela peut refléter des comportements financiers différents selon la profession.



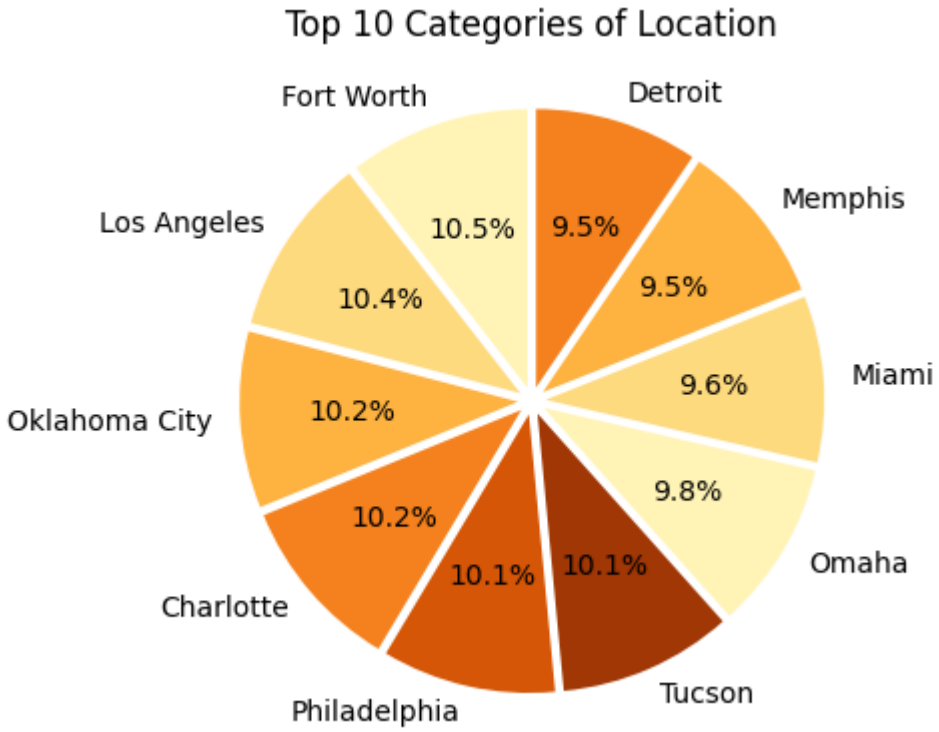


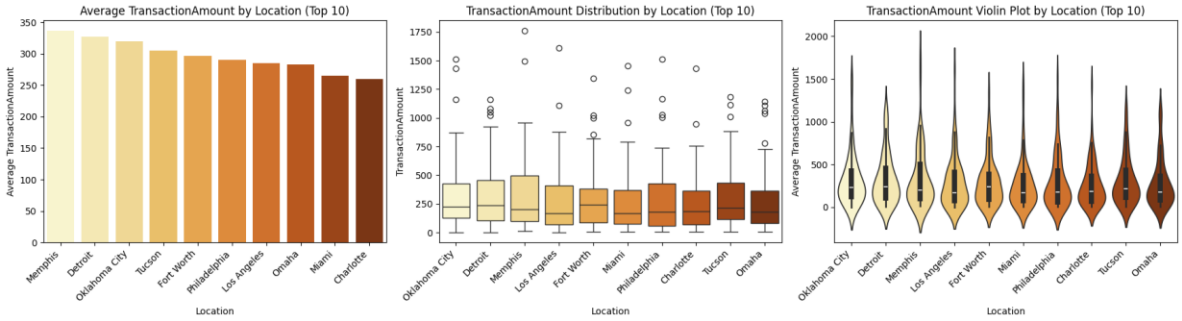


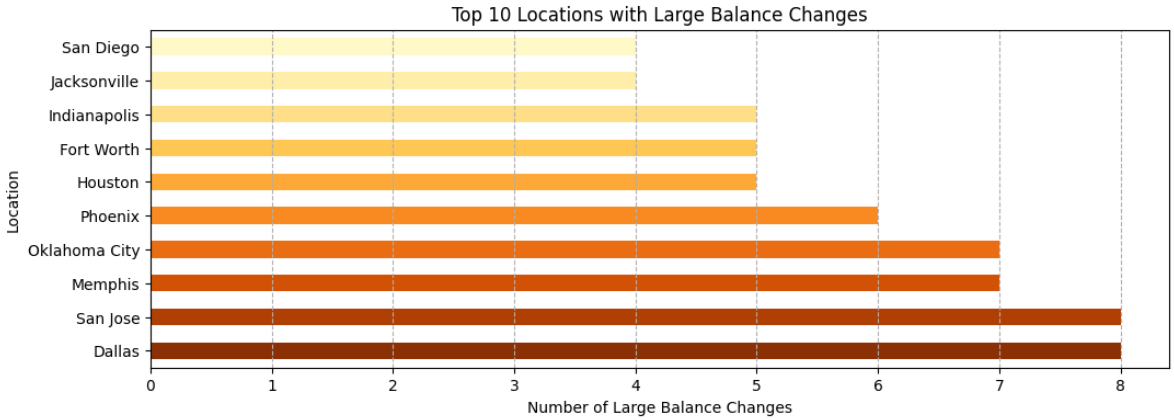
3.6 Analyse géographique des transactions

- Les transactions proviennent de **43 localisations différentes**.
- Les **10 villes principales** concentrent une part importante du volume total, sans qu'une seule ne domine largement.
- La ville **Fort Worth** est la plus active avec **70 transactions**.

Cette diversité géographique permet d'analyser les comportements selon les régions et de détecter d'éventuelles zones à risque.


Top 10 Categories of Location
Fort Worth
Detroit
Memphis
Los Angeles
Miami
Oklahoma City
Omaha
Charlotte
Tucson
Philadelphia
10.5%
9.5%
9.5%
10.4%
9.6%
10.2%
9.8%
10.2%
10.1%
10.1%

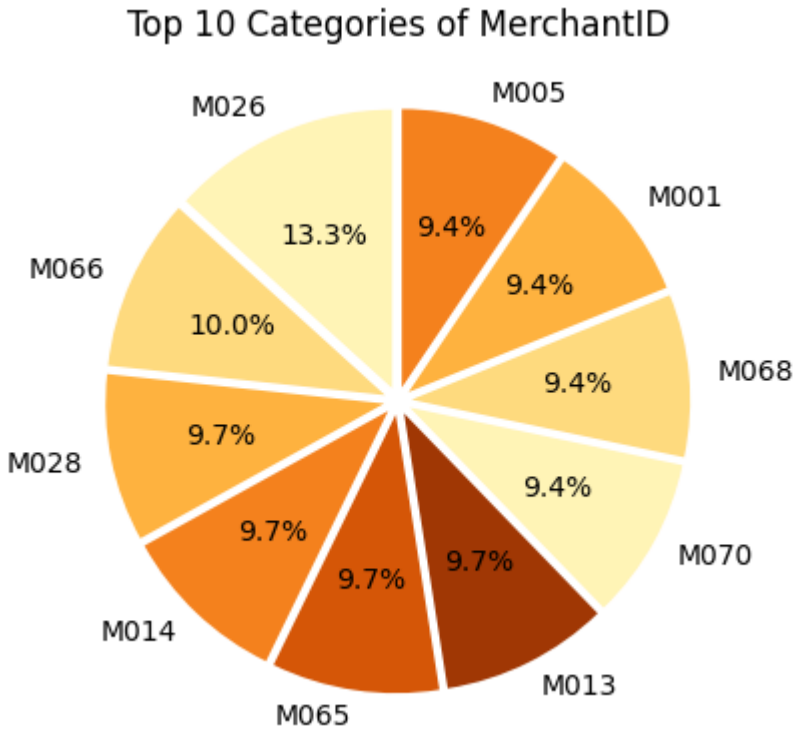

Average TransactionAmount by Location (Top 10)
Average TransactionAmount
Location
TransactionAmount Distribution by Location (Top 10)
TransactionAmount
Location
TransactionAmount Violin Plot by Location (Top 10)
Average TransactionAmount
Location

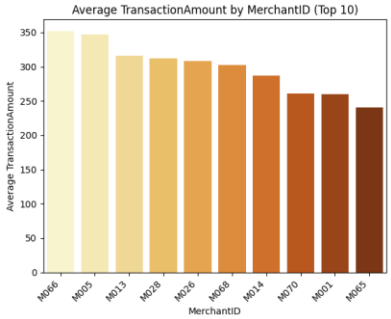

Top 10 Locations with Large Balance Changes
San Diego
Jacksonville
Indianapolis
Fort Worth
Houston
Phoenix
Oklahoma City
Memphis
San Jose
Dallas
Location
Number of Large Balance Changes

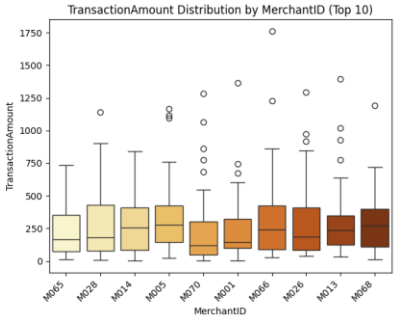
3.7 **Analyse des commerçants (Merchants)**

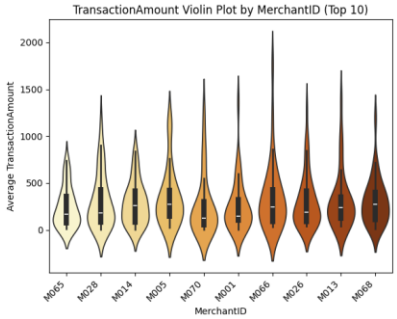
- Le jeu de données comprend **100 commerçants distincts**.
- Le commerçant **M026** se démarque avec **45 transactions**, soit environ **13,3 %** du total.
- Les autres commerçants parmi le top 10 enregistrent entre **9,4 % et 9,7 %** des transactions.

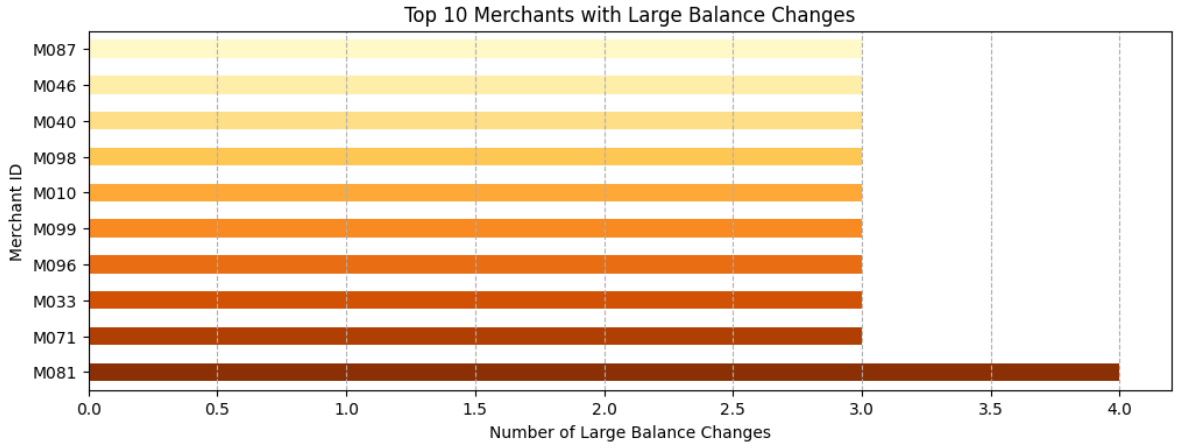
Le volume des transactions reste bien réparti, ce qui reflète un environnement commercial varié.

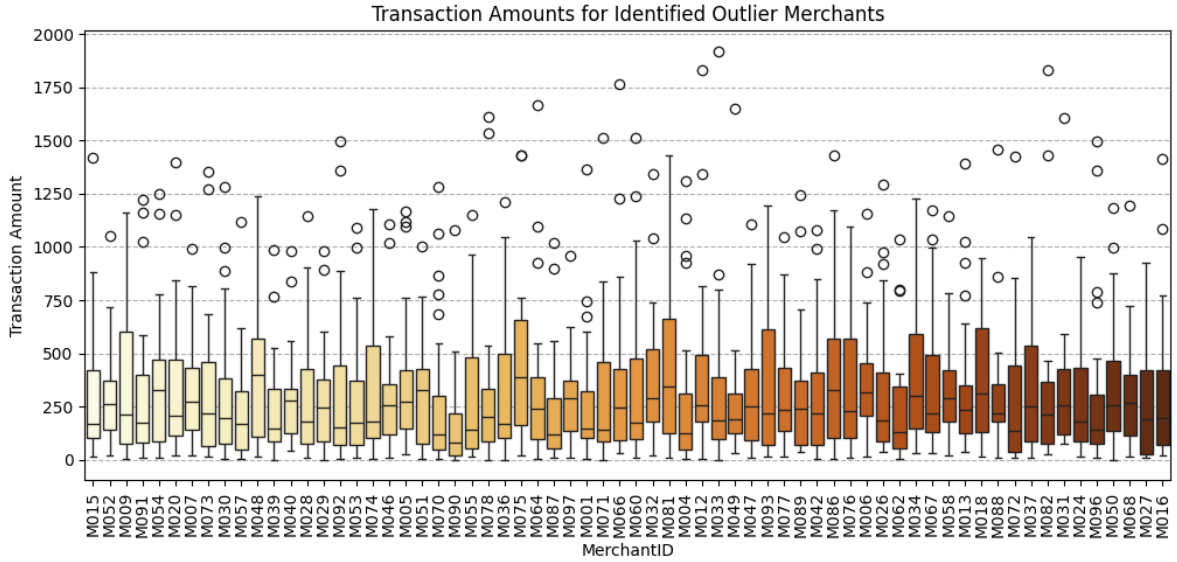












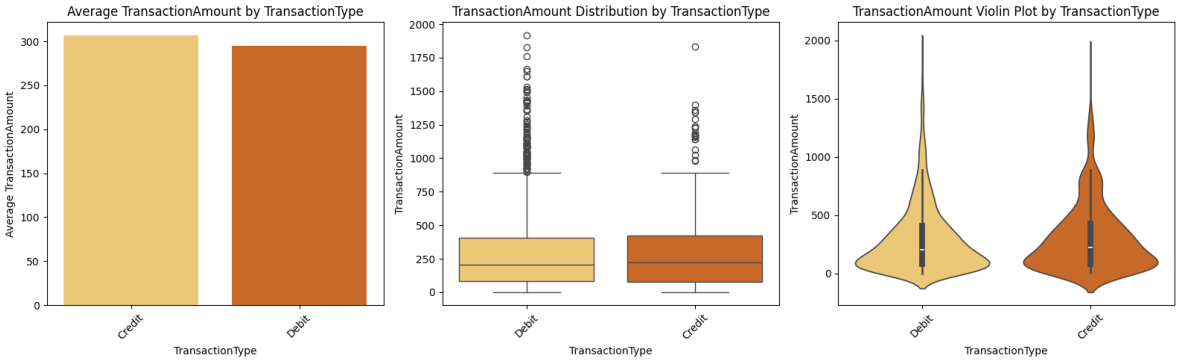
3.8 Comparaison du montant moyen selon le type de transaction

Les analyses graphiques montrent que :

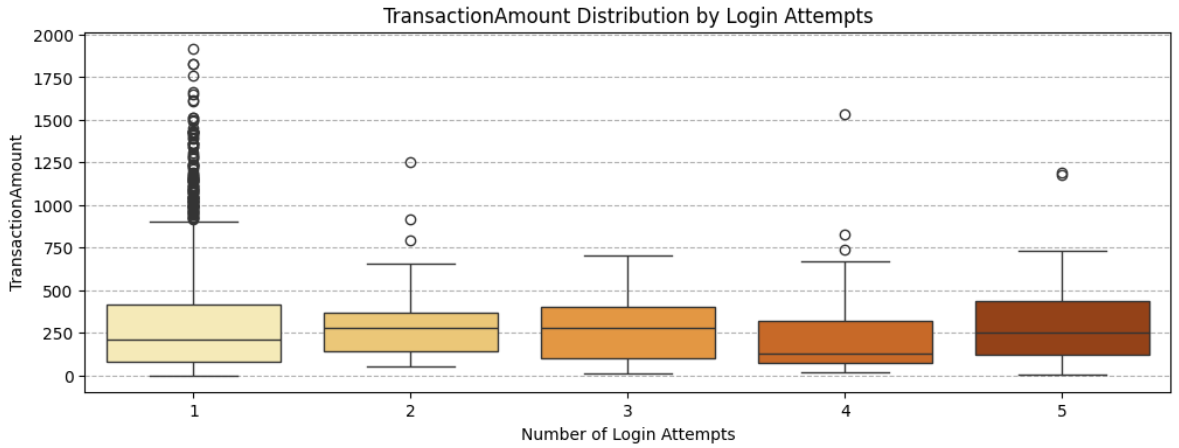
- Les **transactions "Credit"** ont un **montant moyen plus élevé** que les **transactions "Debit"**.
- Les graphiques en violon confirment une **distribution plus étendue** des montants pour les transactions "Credit".

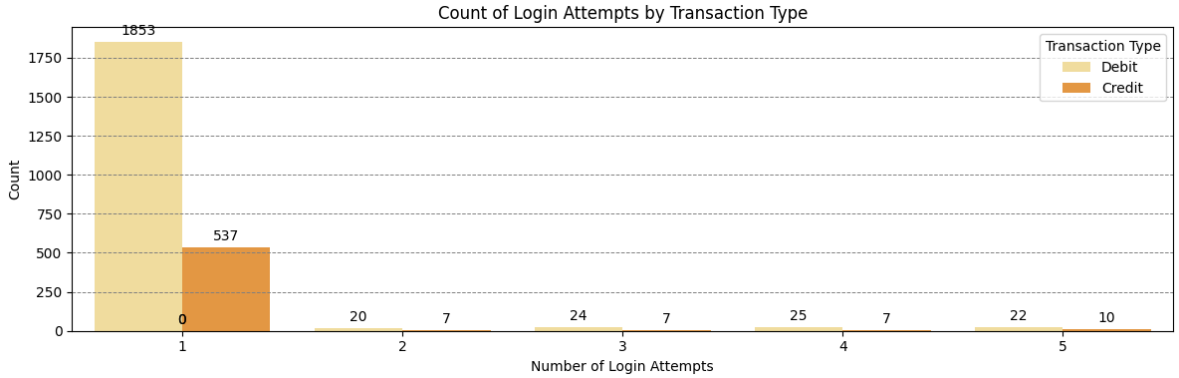
Implications possibles :

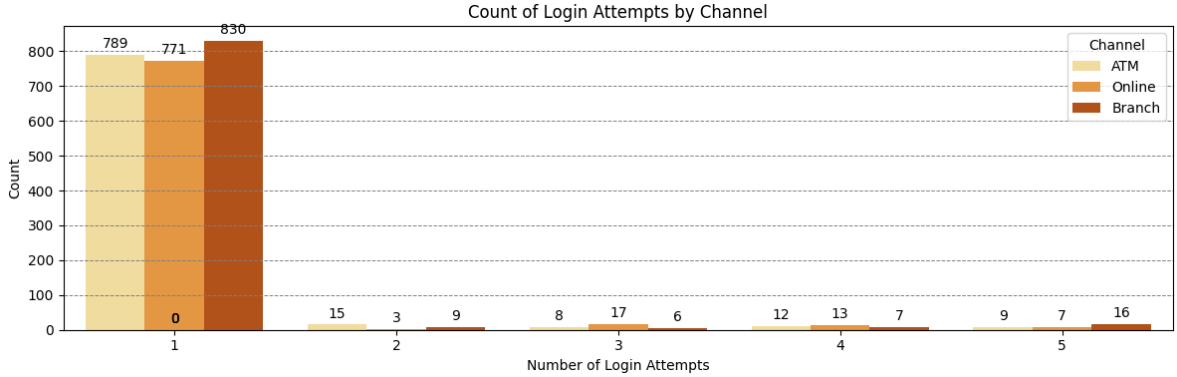
- Les transactions "Credit" correspondent souvent à des **achats plus importants ou à des dépenses exceptionnelles**.
- Ces transactions peuvent présenter un **risque plus élevé de fraude**, en raison de leur valeur plus importante.

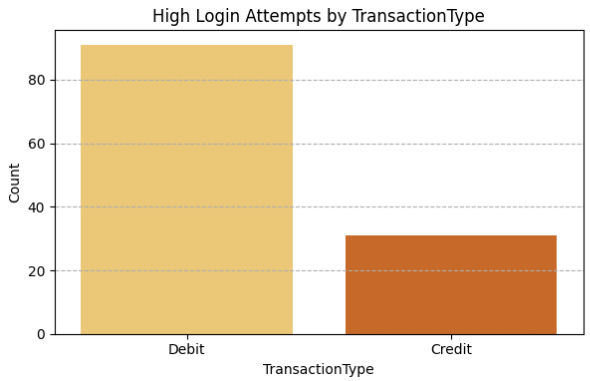


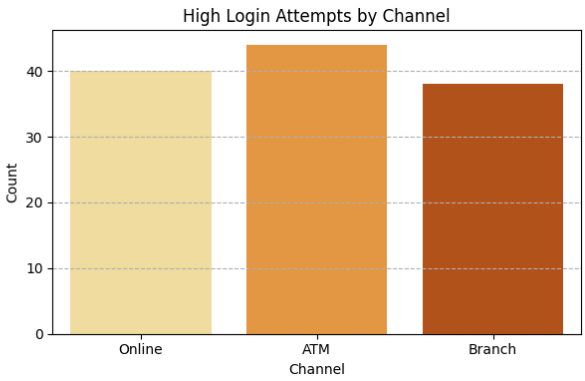
3.9 Quelle est la relation entre le nombre de tentatives de connexion et la répartition des montants des transactions ?

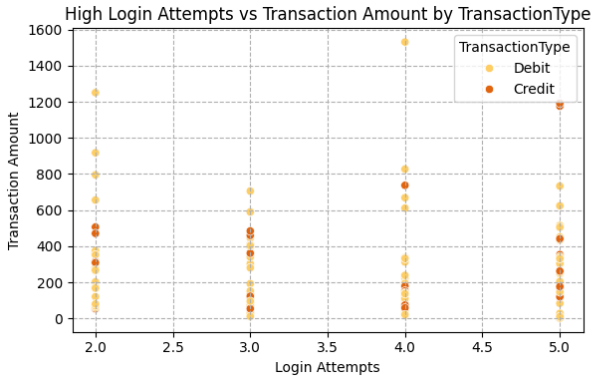


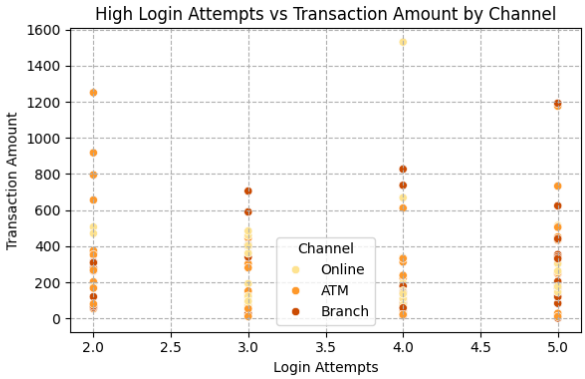










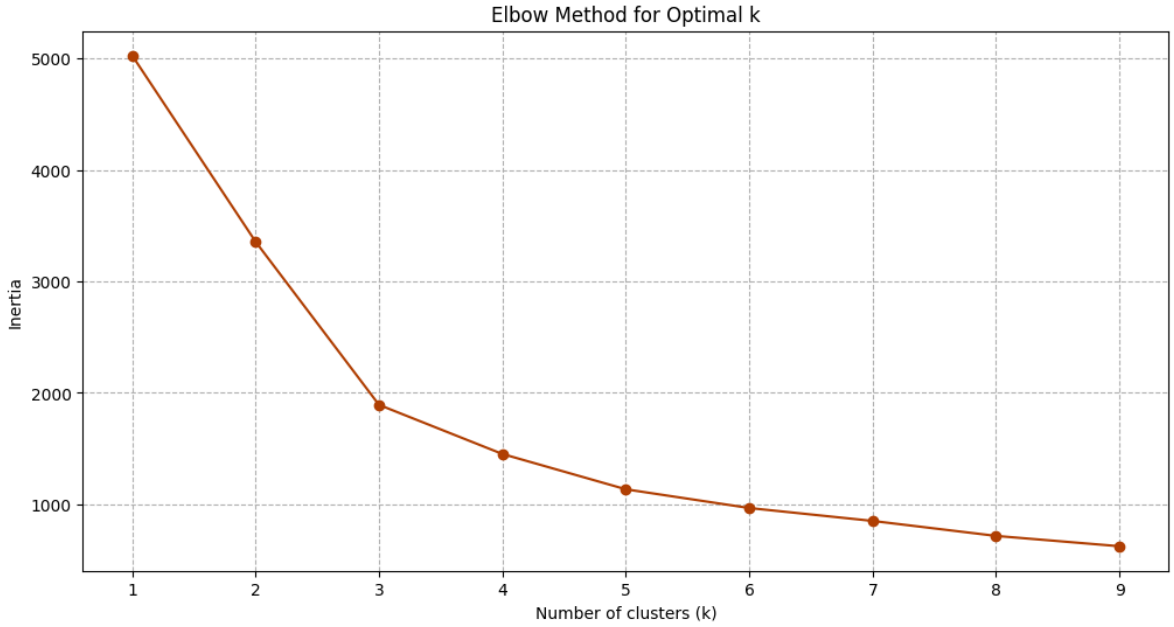


4 Analyse des clusters : Une segmentation claire et détection des fraudes

L'analyse des clusters permet de regrouper les morceaux en groupes homogènes basés sur leurs caractéristiques. Cette segmentation a été réalisée à l'aide de l'algorithme K-Means, DBSCAN et GaussianMixture après une étape de normalisation des données via la méthode StandardScaler pour garantir une échelle uniforme des caractéristiques. Ces méthodologies ont permis de structurer l'ensemble des morceaux en clusters distincts, reflétant des profils variés. L'interprétation détaillée des clusters est présentée ci-dessous.

4.1 Choix du nombre de clusters

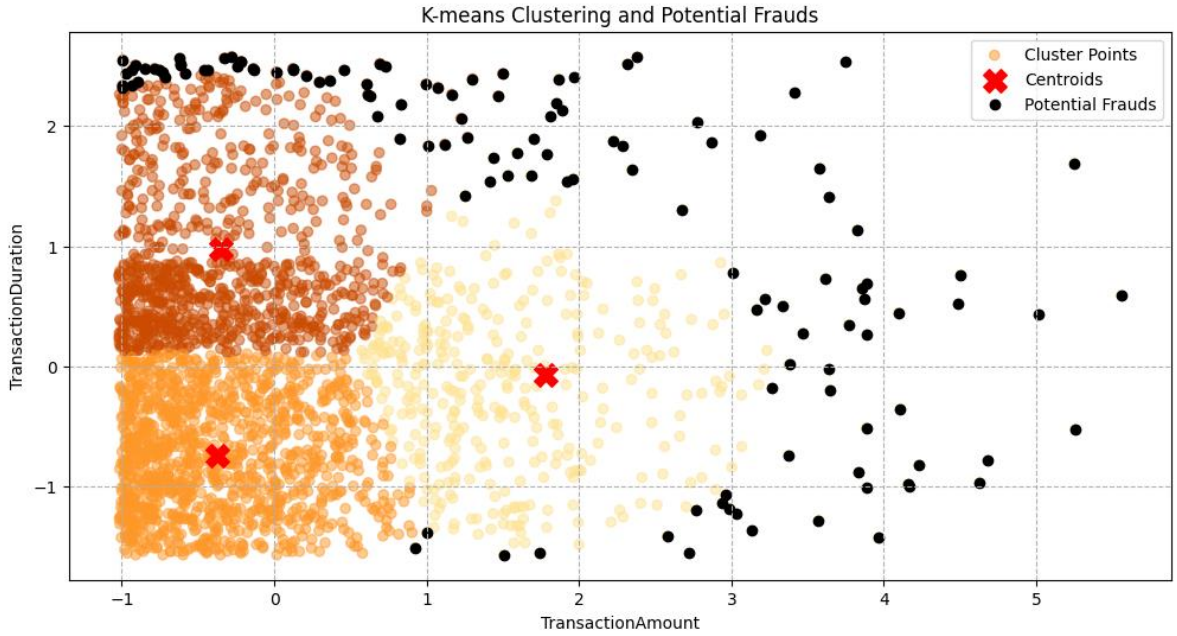
Le graphique du coude (« elbow method ») montre que le point optimal se situe autour de **k = 3**. Ainsi, le modèle K-means avec **3 clusters** semble être un bon choix pour regrouper les transactions.



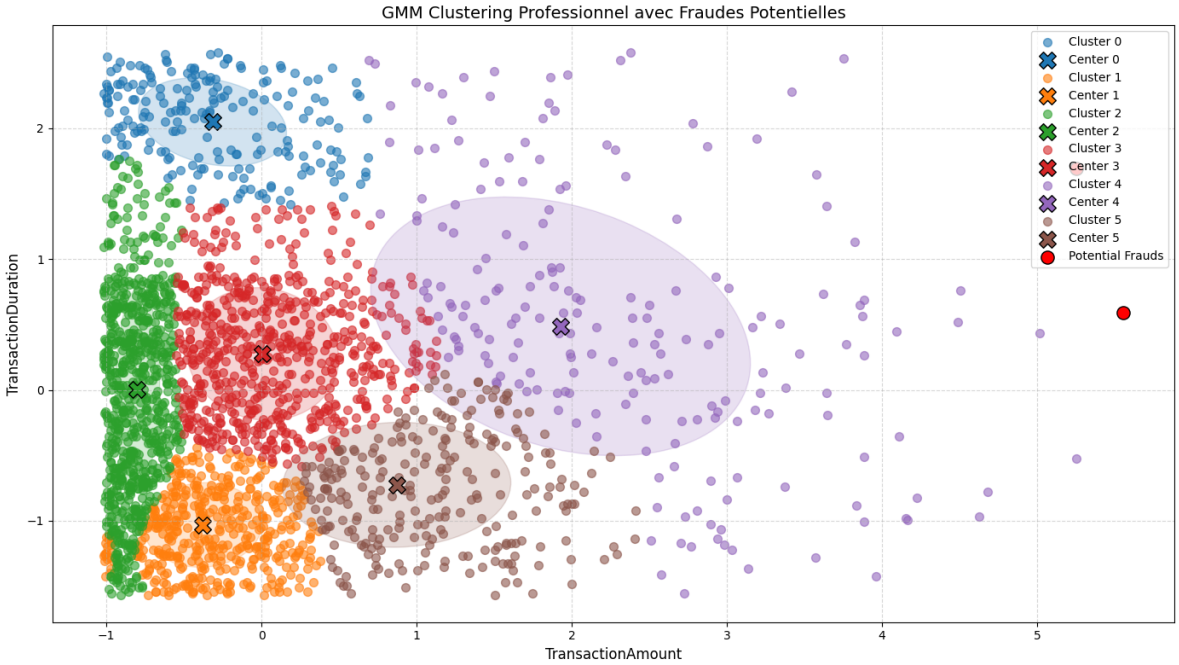
4.2 K-means Clustering

- **Formation des clusters :**
  Le modèle K-means a formé **trois clusters distincts**, bien séparés, ce qui montre une bonne segmentation des transactions similaires.
- **Transactions suspectes :**
  Les **points noirs** représentent les transactions considérées comme **potentiellement frauduleuses**, car elles sont éloignées des centres des clusters.
- **Caractéristiques des clusters :**
  - **Cluster 1 (Jaune) :** transactions normales, fréquentes.
  - **Cluster 2 (Orange) :** transactions de montants et durées moyennes.
  - **Cluster 3 (Marron) :** transactions avec montants et/ou durées plus élevées.
- **Seuil de détection :**
  Le seuil est fixé au **95e percentile** de la distance au centre du cluster. Les points au-delà sont considérés comme suspects.
- **Nombre de fraudes détectées :**
  Environ **126 transactions** ont été identifiées comme potentiellement frauduleuses.

Le modèle K-means permet d'identifier efficacement les transactions qui s'écartent des comportements habituels.



4.3 Gaussian Mixture Clustering

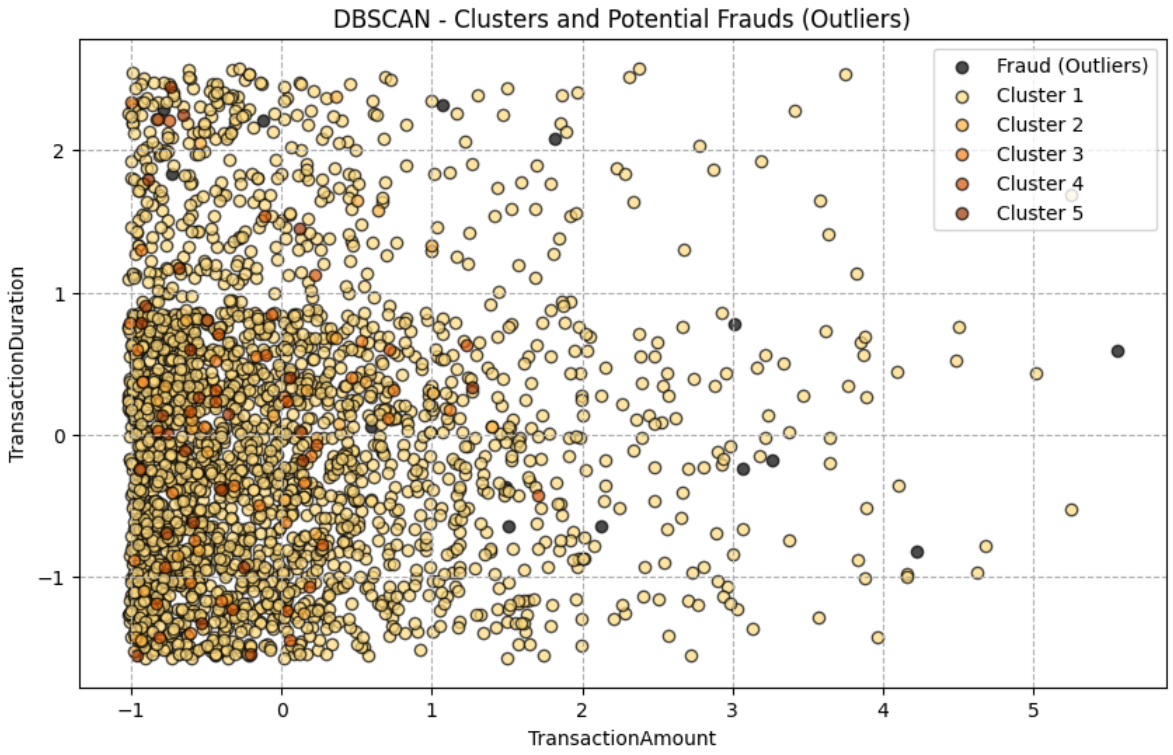


4.4 DBSCAN Clustering

- **Formation des clusters :**
  DBSCAN a détecté plusieurs groupes de transactions, reflétant des comportements similaires.
- **Fraudes potentielles (outliers) :**
  Les **points noirs** représentent les transactions isolées, identifiées comme **frauduleuses**.
  Ces points sont dispersés et ne font partie d'aucun cluster dense.

- **Caractéristiques :**
  - Les variables ont été **standardisées** avant l'analyse pour de meilleures performances.
  - DBSCAN repère les **zones de faible densité** pour détecter les anomalies.
- **Nombre de fraudes détectées :**
  Environ **17 transactions** ont été détectées comme suspectes.

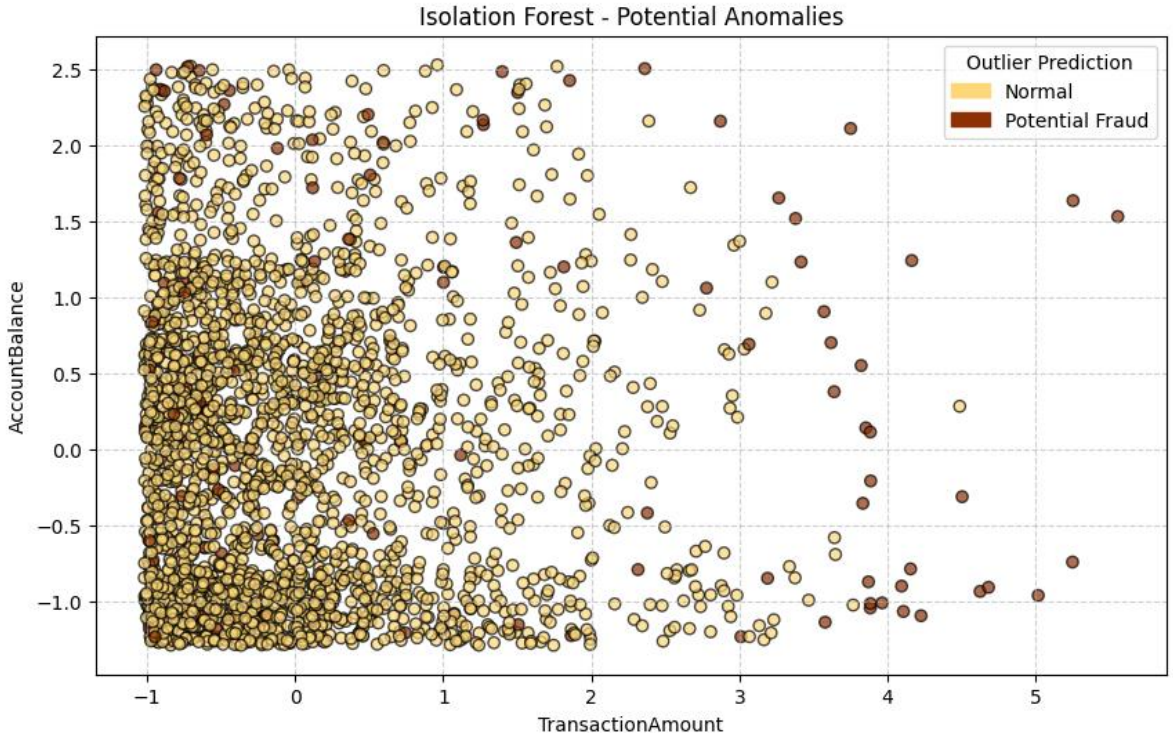
DBSCAN détecte efficacement les transactions anormales grâce à l'analyse de densité locale.



4.5 Isolation Forest

- **Principe :**
  Isolation Forest détecte les anomalies en isolant les points rares dans les données.
- **Résultats :**
  - Les **points sombres** indiquent les transactions suspectes.
  - Les **points clairs** correspondent aux transactions normales.
  - La visualisation montre la répartition selon **TransactionAmount** et **AccountBalance**.
- **Prédiction et score d'anomalie :**
  Le modèle attribue un score d'anomalie à chaque transaction pour déterminer si elle est normale ou non.
- **Nombre de fraudes détectées :**
  Environ **126 transactions** ont été identifiées comme potentiellement frauduleuses.

L'Isolation Forest s'avère très efficace pour repérer les transactions inhabituelles.



5 Conclusion

Ce projet avait pour objectif de **détecter les transactions frauduleuses** à partir d'un jeu de données financières, en combinant des **techniques d'analyse exploratoire** et des **méthodes de Machine Learning non supervisées**.

Dans un premier temps, l'**analyse exploratoire des données (EDA)** a permis de mieux comprendre la structure du dataset et d'identifier certaines relations entre les variables. Par exemple, une corrélation positive a été observée entre le montant des transactions et le solde du compte, tandis que l'âge du client ou le nombre de tentatives de connexion n'avaient pas d'influence notable sur le comportement transactionnel. Ces observations ont permis de poser des bases solides pour la phase de modélisation.

Ensuite, plusieurs **méthodes de clustering et de détection d'anomalies** ont été mises en œuvre afin d'identifier les comportements suspects :

- **K-means** a permis une segmentation claire des transactions en trois groupes distincts, dont un groupe de points éloignés représentant des transactions potentiellement frauduleuses.
- **DBSCAN** a détecté des anomalies basées sur la densité locale, révélant plusieurs transactions isolées.
- **Isolation Forest** s'est montré particulièrement efficace pour isoler les anomalies, confirmant la présence d'un ensemble de transactions atypiques.

Les résultats obtenus montrent que ces approches peuvent **compléter efficacement les systèmes de détection de fraude traditionnels**, en repérant des transactions inhabituelles sans supervision préalable.

En conclusion, ce travail met en évidence l'importance des **méthodes non supervisées** dans la **lutte contre la fraude financière**.