Synthèse

Le « dataset » utilisé pour cette analyse contient **284 807 transactions**, chacune décrite par **31 variables**. Parmi celles-ci, **30 sont de type numérique (float64)** et correspondent aux composantes principales issues d'une transformation PCA (V1 à V28, ainsi que Time et Amount). La dernière colonne, **Class**, est un entier (0 ou 1) représentant la nature de la transaction (normale ou frauduleuse). On observe également 284,315 transactions normales contre seulement 492 frauduleuses. Cela représente **moins de 0,18 % de fraudes**, ce qui souligne un **fort déséquilibre** qui peut biaiser l'entraînement d'un modèle de classification. Toutes les variables du dataset sont **complètes**, c'est-à-dire qu'aucune valeur manquante n'a été détectée.

Observations principales :

- Les variables V1 à V28 sont issues d'une **réduction par ACP (Analyse en Composantes Principales)** → elles sont anonymisées, ce qui limite leur interprétation directe.
- La variable Amount (montant) a une moyenne de **88.35 €**, mais un **écart-type élevé (250 €) → distribution très asymétrique à droite**.
- Le boxplot du montant des transactions met en évidence un grand nombre de valeurs aberrantes. La plupart des transactions ont un montant inférieur à 100 €, mais certaines atteignent plus de 25 000 €(outliers). Cette hétérogénéité justifie une transformation du montant (comme la transformation logarithmique) pour stabiliser la variance avant l'entraînement d'un modèle.
- La variable Time indique le nombre de secondes écoulées depuis la première transaction. La distribution temporelle montre une structure bimodale : deux pics d'activité sont observés, séparés par une période creuse. Cela peut correspondre à des plages horaires distinctes (par exemple, jour et nuit), ce qui suggère que le moment de la transaction pourrait être une variable pertinente pour détecter des comportements inhabituels.
- La variable Class est fortement déséquilibrée : seulement 0,17 % des transactions sont frauduleuses. Ce déséquilibre extrême implique que des métriques classiques comme l'exactitude (accuracy) ne sont pas appropriées pour évaluer la performance d'un modèle. Des techniques telles que **le suréchantillonnage** ou **le sous-échantillonnage** des données majoritaires seront également nécessaires pour rééquilibrer les classes lors de l'entraînement.

Conclusion :

Cette analyse exploratoire met en lumière plusieurs caractéristiques essentielles du jeu de données : un fort déséquilibre de classes, une distribution très asymétrique des montants, ainsi qu'une structure temporelle intéressante. Ces éléments orientent les choix à venir en termes de prétraitement (normalisation, transformation logarithmique) et de stratégie de modélisation (techniques de rééquilibrage, algorithmes robustes aux données déséquilibrées, validation croisée adaptée).

La phase suivante consistera à **préparer les données** (scaling, réduction du déséquilibre) avant de tester différents modèles de classification adaptés à la détection de fraudes.