IUT de Paris - Rives de Seine
Université Paris Cité

jupyter

RAPPORT DU PROJET : PROGRAMMATION STAISTIQUE

programmation statistique avec jupyter

Chakir Hajar
Dris Ines
BUT 2 SD/EMS

METHODOLOGIE DU RAPPORT

INTRODUCTION

Ce projet vise à analyser les données de crédits provenant d'une banque afin d'évaluer le profil des clients et de déterminer s'ils peuvent être classés comme « bon client » ou « mauvais client ». Les données comprennent 1000 observations et 20 variables, couvrant des aspects financiers, personnels et bancaires des clients.

La variable cible est binaire :

- 1 : Bon client
- 2 : Mauvais client

L'objectif principal est de fournir une analyse exploratoire et descriptive des données à travers des visualisations, des statistiques et des méthodes multidimensionnelles comme l'analyse en composantes principales (ACP) et les algorithmes de clustering. Ces analyses permettront de mieux comprendre les relations entre les variables et d'identifier les patrons sous-jacents parmi les clients.

Pour mener à bien cette analyse, plusieurs outils et bibliothèques Python ont été utilisés :

- Pandas : gestion et manipulation des données
- NumPy : calculs numériques
- SciPy : outils statistiques
- Matplotlib : visualisations de données
- Scikit-learn : méthodes d'apprentissage automatique et analyses multidimensionnelles

PRÉSENTATION DES DONNÉES

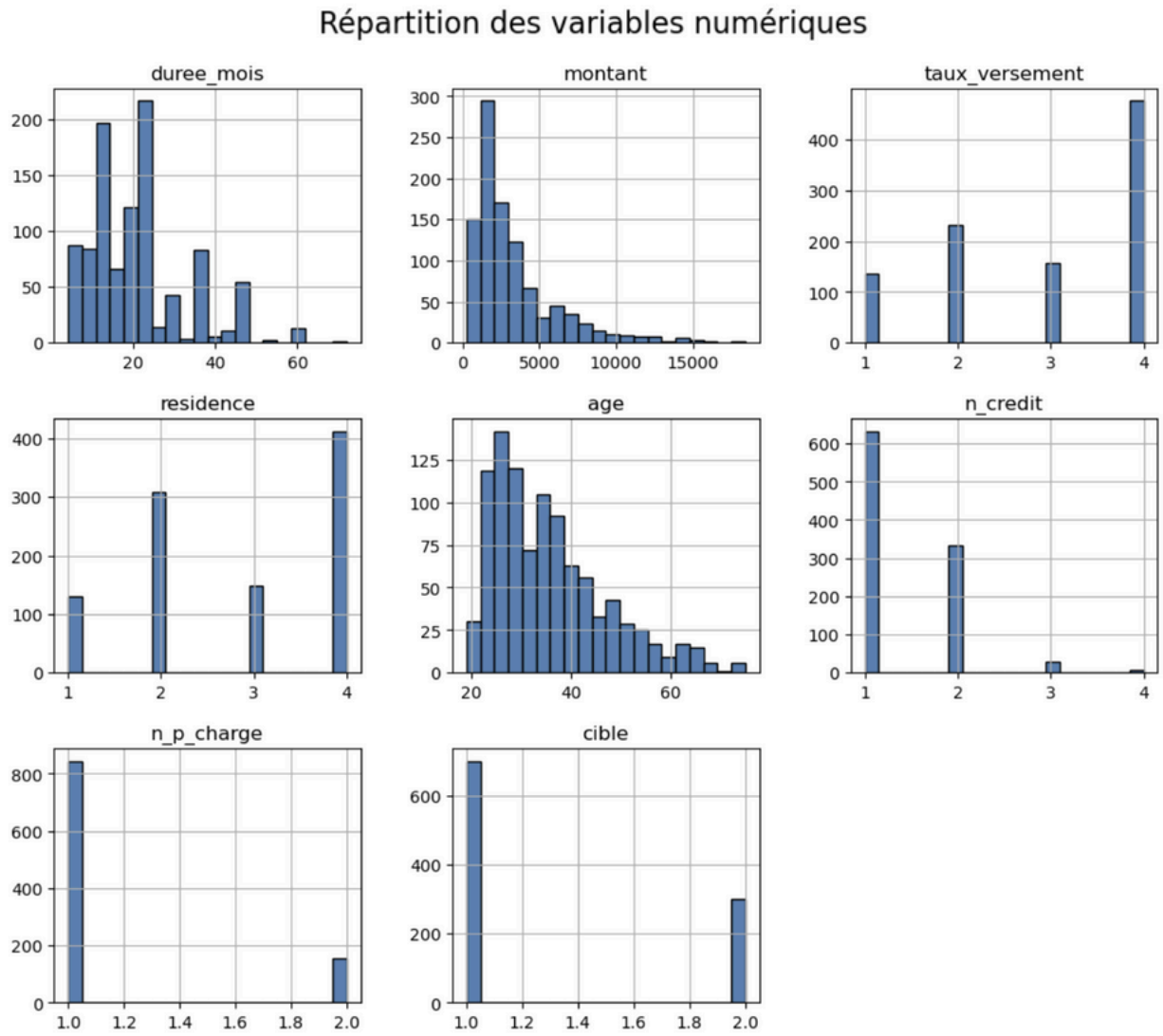
Variables Numériques	Description
duree_mois	Durée du crédit en mois
montant	Montant du crédit en euros
taux_versement	Taux de versement en pourcentage du revenu disponible
residence	Durée de résidence actuelle (en années)
age	Âge en années
n_credit	Nombre de crédits existants dans cette banque
n_p_charge	Nombre de personnes à charge

PRÉSENTATION DES DONNÉES

Variables catégorielles	Description
statut_compte	Statut du compte courant existant
historique	Historique de crédit
objectif	Objectif du crédit
epargne	Compte d'épargne ou bons
emploi	Emploi actuel depuis combien de temps
statut_sexe	Statut personnel et sexe
autre_debiteurs	Autres débiteurs ou garants
propriete	Type de propriété détenue
plan_versement	Autres plans de versement
logement	Type de logement (propriétaire, locataire, etc.)
travail	Type de travail
tel	Possession d'un téléphone (oui/non)
trav_etranger	Statut de travailleur étranger (oui/non)
variable cible	Indique si le client est bon (1) ou mauvais (2) pour le crédit

ANALYSE EXPLORATOIRE DES DONNÉES

I/Analyse Univariée



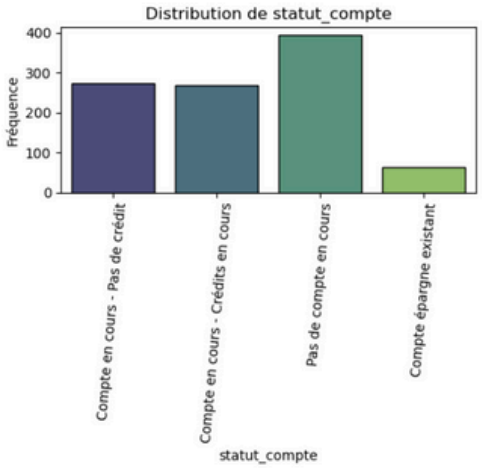
Les graphiques univariés permettent de visualiser la répartition de chaque variable numérique individuellement. Ils sont utiles pour comprendre les **tendances générales** et identifier des **concentrations** ou des **anomalies** dans les données. Par exemple, ils nous aident à voir si une variable est uniformément répartie ou si elle présente des pics ou des creux significatifs.

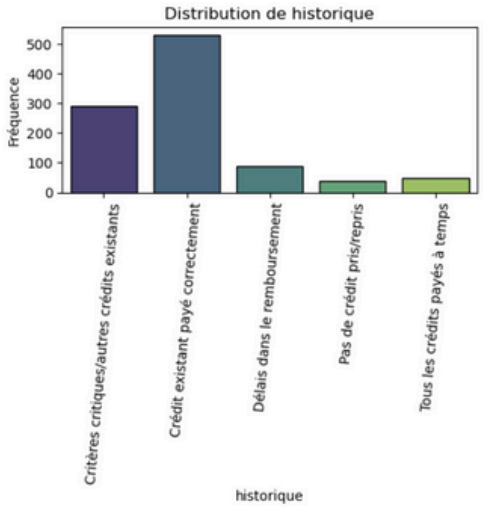
Prenons le graphique de l'âge par exemple, On observe que la majorité des clients ont entre 20 et 40 ans, avec un pic autour de 30 ans. Cela montre que les emprunteurs sont principalement de jeunes adultes ou des personnes d'âge moyen, ce qui pourrait refléter des besoins financiers ou une capacité d'emprunt spécifique à cette tranche d'âge.

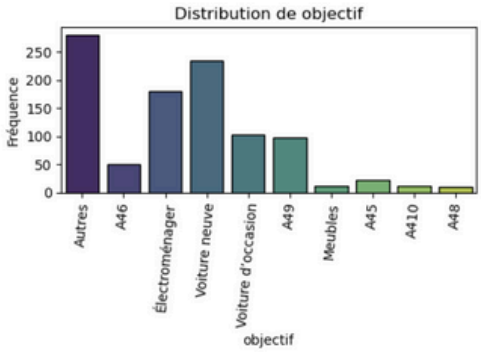
En résumé, ces graphiques sont une première étape essentielle pour **mieux comprendre les données** avant d'approfondir les analyses.

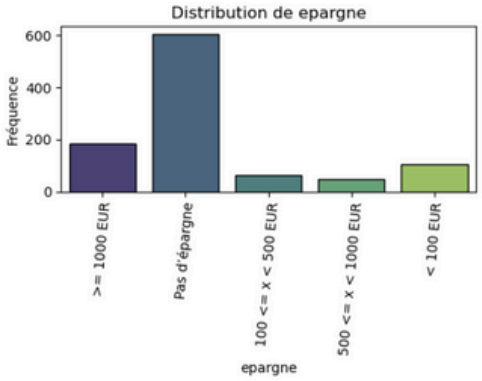
I/Analyse Univariée

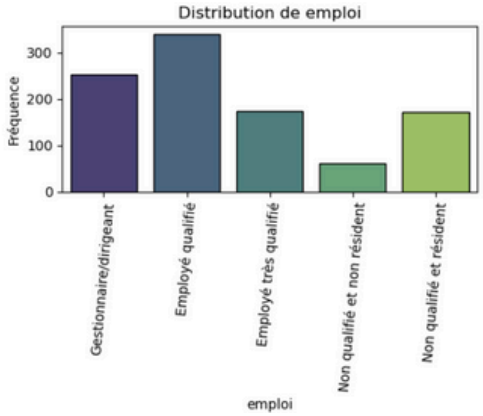
La répartition de chaque variable catégorielles

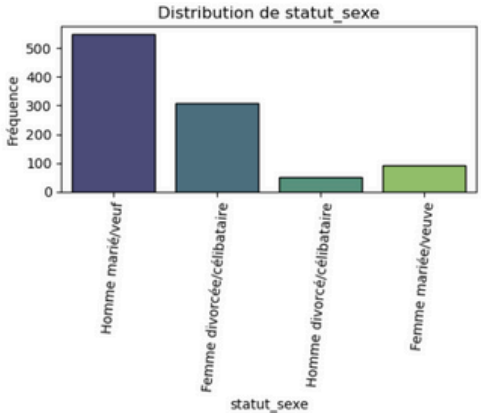


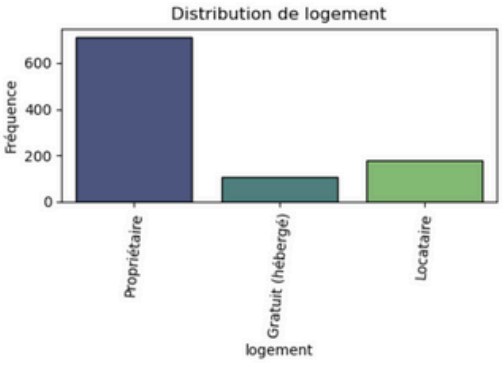


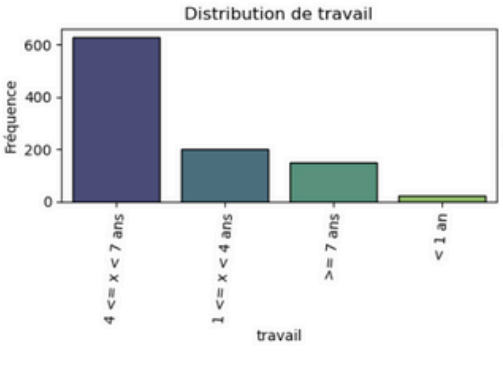


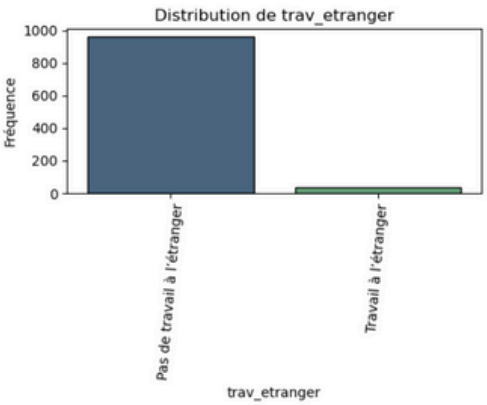


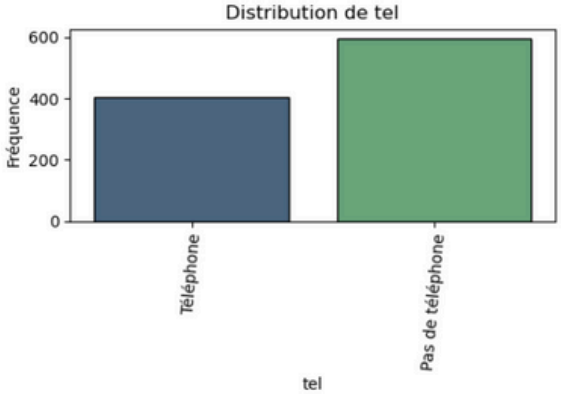


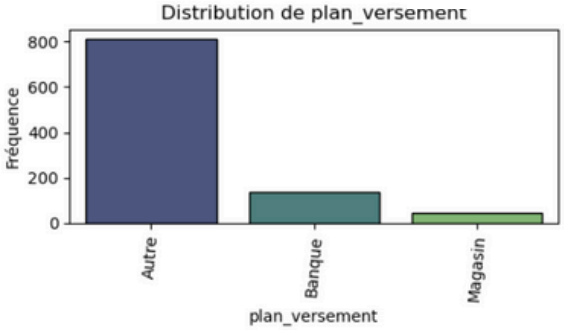












ANALYSES EXPLORATOIRE DES DONNÉES

I/Analyse Univariée

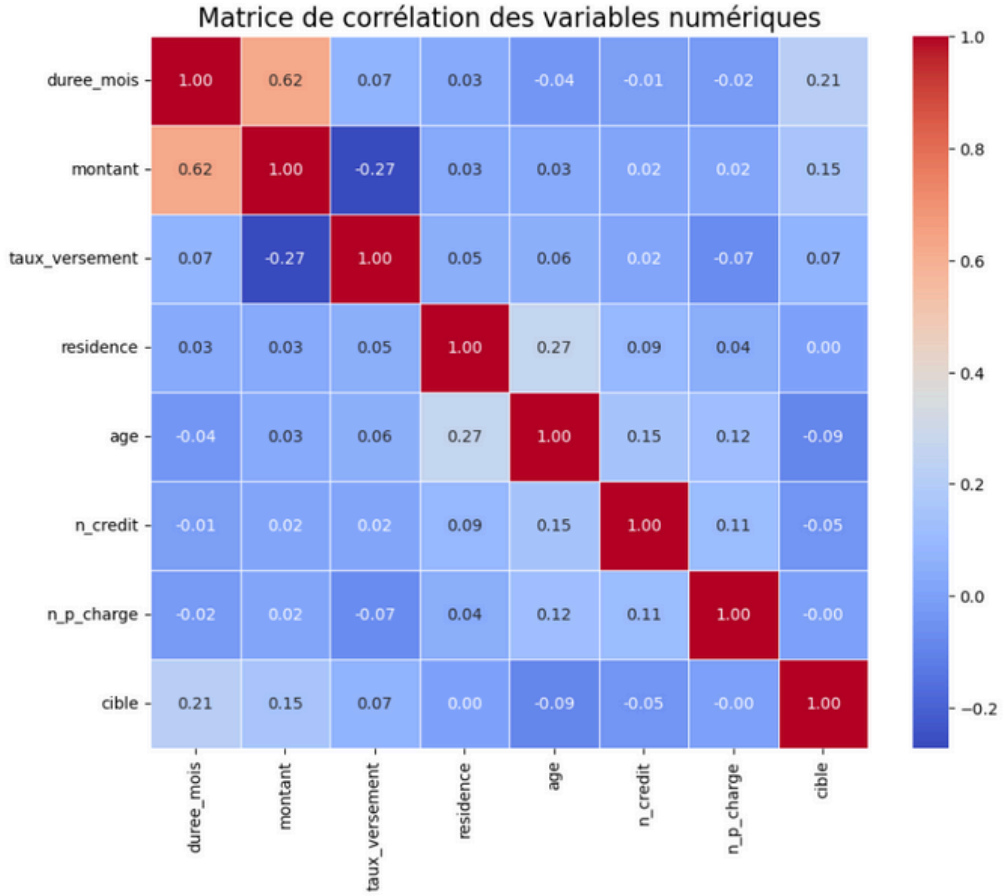
Les graphiques univariés des variables catégorielles permettent de visualiser la fréquence des différentes catégories présentes dans les données. Ils sont utiles pour comprendre la **répartition des caractéristiques** des clients, repérer les **catégories dominantes**, et **détecter des anomalies** ou **déséquilibres** dans les données. Ces informations sont essentielles pour interpréter les comportements et prioriser des analyses plus approfondies.

Par exemple, dans le graphique de la variable **“statut_sexe”**, on observe que les hommes mariés ou veufs représentent une majorité significative, tandis que les femmes mariées ou veuves sont minoritaires.
Concernant la variable **“epargne”**, plus de 400 clients n’ont pas d’épargne, et moins de 100 ont une épargne supérieure à 1000 EUR, ce qui montre une forte disparité dans la capacité d’épargne.
Enfin, pour la variable **“travail”**, la majorité des clients ont une ancienneté inférieure à 4 ans, avec un pic entre 1 et 4 ans, tandis que ceux ayant une ancienneté supérieure à 7 ans sont peu nombreux.

En conclusion, ces graphiques permettent d’identifier les tendances générales et les disparités parmi les différentes catégories. Cela aide à mieux segmenter les clients et à adapter les stratégies d’analyse ou de prise de décision en fonction des caractéristiques observées.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



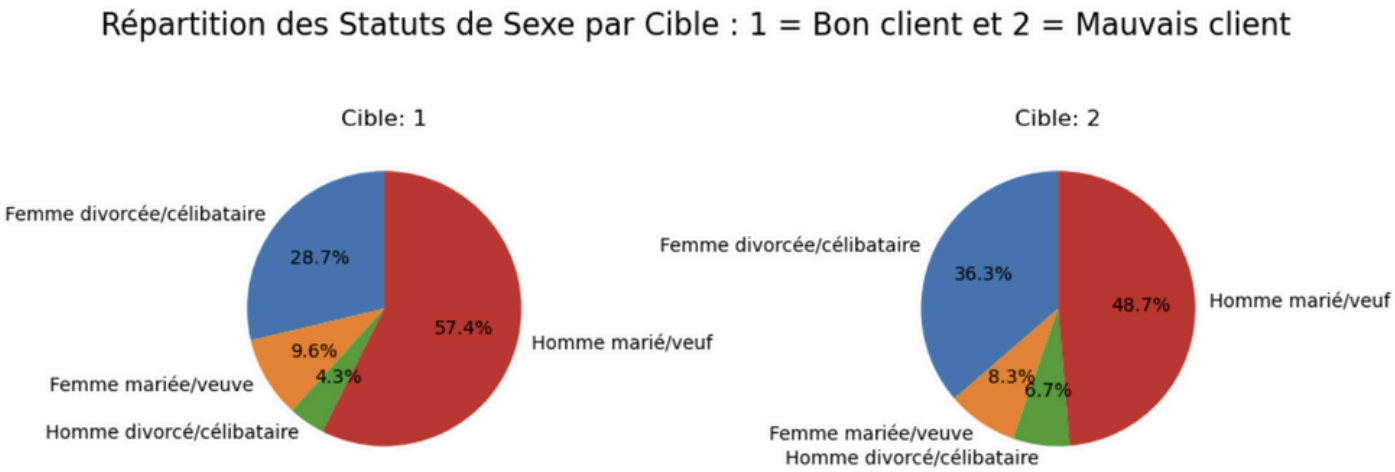
Cette matrice de corrélation permet d'évaluer les relations linéaires entre les variables numériques afin d'identifier les dépendances et les interactions possibles. Elle sert à mieux comprendre comment les variables influencent la variable cible, ici la classification des bons et mauvais clients.

La variable durée_mois montre une corrélation modérée avec montant à 0,62, ce qui indique que des crédits plus longs sont souvent associés à des montants plus élevés. Aucune des variables n'a de corrélation forte avec la cible, mais durée_mois et montant montrent des relations légèrement positives à 0,21 et 0,15 respectivement.
Les variables comme résidence, âge et n_p_charge ont des corrélations très faibles avec la cible, ce qui suggère un impact limité sur la classification des clients. Une corrélation négative faible entre taux_versement et montant à -0,27 indique que des taux de versement plus élevés sont liés à des montants de crédit plus faibles.

En résumé, cette matrice est utile pour identifier les relations significatives entre les variables et pour cibler celles qui méritent une attention particulière dans les analyses futures. Les faibles corrélations avec la cible suggèrent que la classification des clients est probablement influencée par une combinaison de plusieurs facteurs, justifiant une exploration plus approfondie avec des analyses multivariées.

ANALYSES EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



Ce graphique permet de mettre en lumière la répartition des clients selon leur statut marital et leur sexe, en fonction de leur classification en bons ou mauvais clients.
Il est utile pour identifier les profils dominants dans chaque catégorie cible, ce qui peut aider à adapter les stratégies d'attribution de crédit en fonction des caractéristiques des clients.

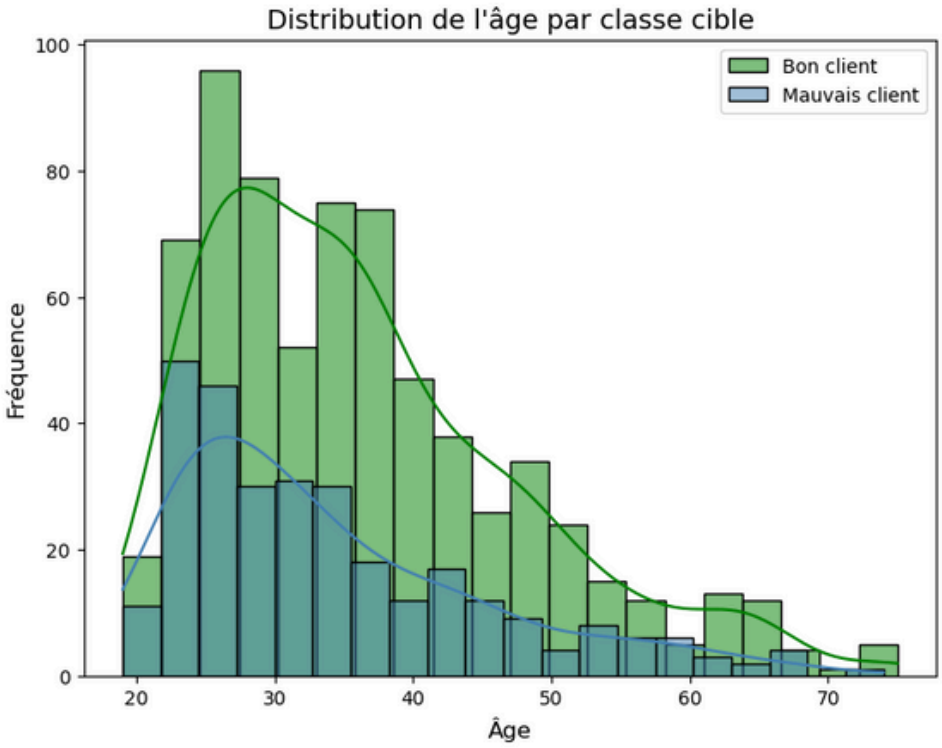
Pour la Cible 1 (bons clients), 57,4 % sont des hommes mariés ou veufs, suivis des femmes divorcées ou célibataires (28,7 %). Cette prédominance des hommes mariés ou veufs dans l'ensemble des données peut toutefois biaiser les résultats, car leur proportion est nettement plus élevée dans la population globale analysée.

À l'inverse, dans la Cible 2 (mauvais clients), les hommes mariés ou veufs représentent 48,7 %, tandis que la proportion de femmes divorcées ou célibataires augmente à 36,3 %.

En conclusion, ce graphique montre que les hommes mariés ou veufs sont majoritairement bons clients, tandis que les femmes divorcées ou célibataires sont plus souvent associées aux mauvais clients, ce qui peut révéler des différences dans les comportements financiers.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



Ce graphique montre la répartition de l'âge des clients en fonction de leur classification en bons ou mauvais clients. Il permet d'observer les tranches d'âge les plus représentées dans chaque catégorie cible et d'identifier des différences intéressantes entre les deux groupes.

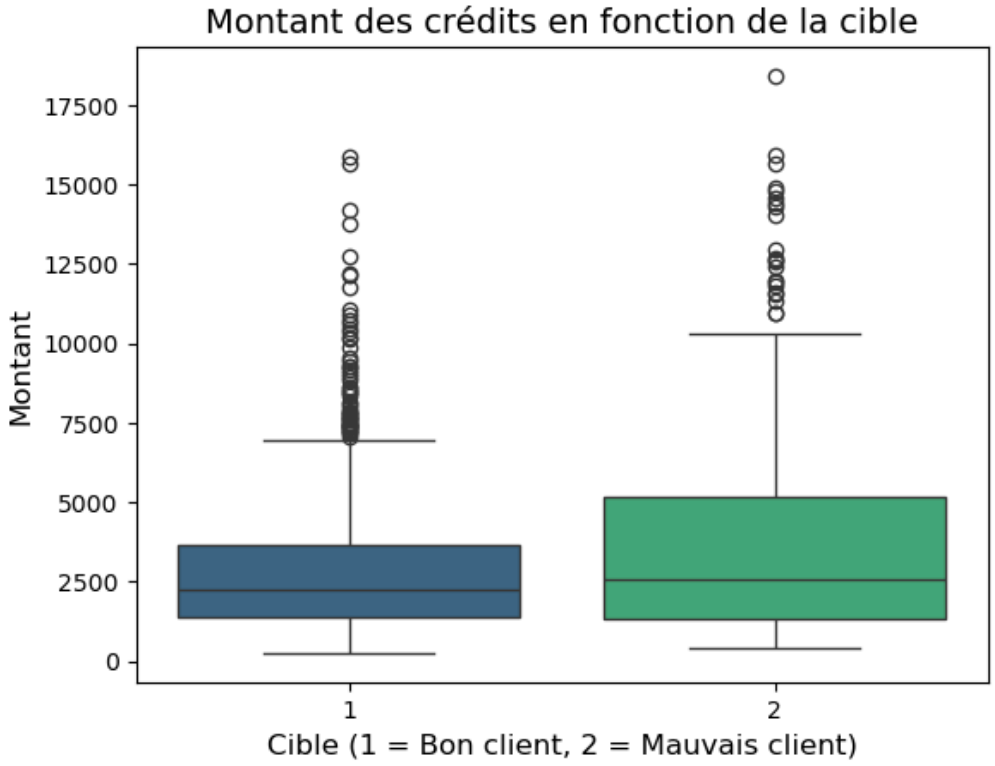
Pour les bons clients, on observe une concentration importante dans la tranche d'âge des 20-40 ans, avec un pic autour de 30 ans, ce qui montre que les jeunes adultes sont majoritairement considérés comme fiables pour les crédits. En revanche, pour les mauvais clients, bien que cette tranche d'âge soit également représentée, la fréquence est globalement plus faible et les mauvais clients sont plus dispersés sur les âges supérieurs à 40 ans.

On peut également noter que la proportion de mauvais clients augmente légèrement dans les âges plus avancés, tandis que celle des bons clients diminue. Cela pourrait indiquer un lien entre l'âge, le comportement financier, et la capacité de remboursement.

Ce graphique met ainsi en évidence que l'âge est un facteur qui semble jouer un rôle dans la distinction entre bons et mauvais clients, en particulier pour les jeunes adultes et les personnes plus âgées.

ANALYSES EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



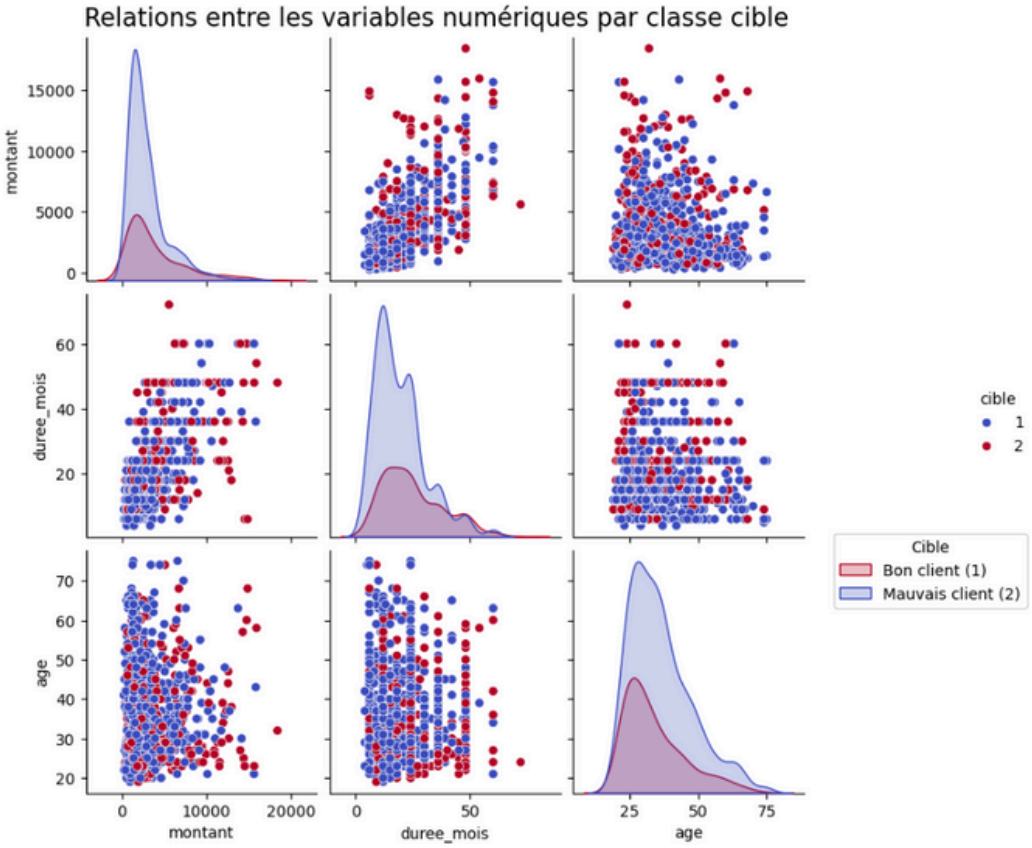
Ce graphique, sous forme de boîtes à moustaches, montre la distribution des montants des crédits en fonction de la classification **des bons clients** (Cible 1) et **mauvais** (Cible 2). Il est utile pour identifier les différences dans les montants empruntés selon la performance des clients, tout en mettant en évidence les valeurs extrêmes.

Pour les bons clients (Cible 1), la médiane du montant des crédits est autour de 2500, avec une concentration de données entre 1000 et 4000. Cependant, on observe quelques valeurs extrêmes dépassant 10 000. En revanche, pour les mauvais clients (Cible 2), la médiane est légèrement plus élevée, et les crédits sont plus dispersés, avec un écart interquartile plus large, ce qui indique une plus grande variabilité.
On remarque également plus de crédits importants parmi les mauvais clients, certains montant dépassant 15 000.

En conclusion, ce graphique met en lumière que les mauvais clients ont tendance à emprunter des montants plus élevés et plus dispersés, ce qui peut représenter un facteur de risque supplémentaire pour les institutions financières. Cela souligne l'importance d'analyser les montants des crédits pour mieux comprendre le comportement des clients.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



Ce graphique, met en relation les variables numériques montant, durée_mois, et âge selon la classification des clients en bons (Cible 1) et mauvais (Cible 2). Il est particulièrement utile pour observer les relations entre ces variables, identifier des tendances générales et repérer les différences entre les deux groupes. Les points représentent les observations individuelles (les clients), et leur position reflète les valeurs associées à chaque variable.

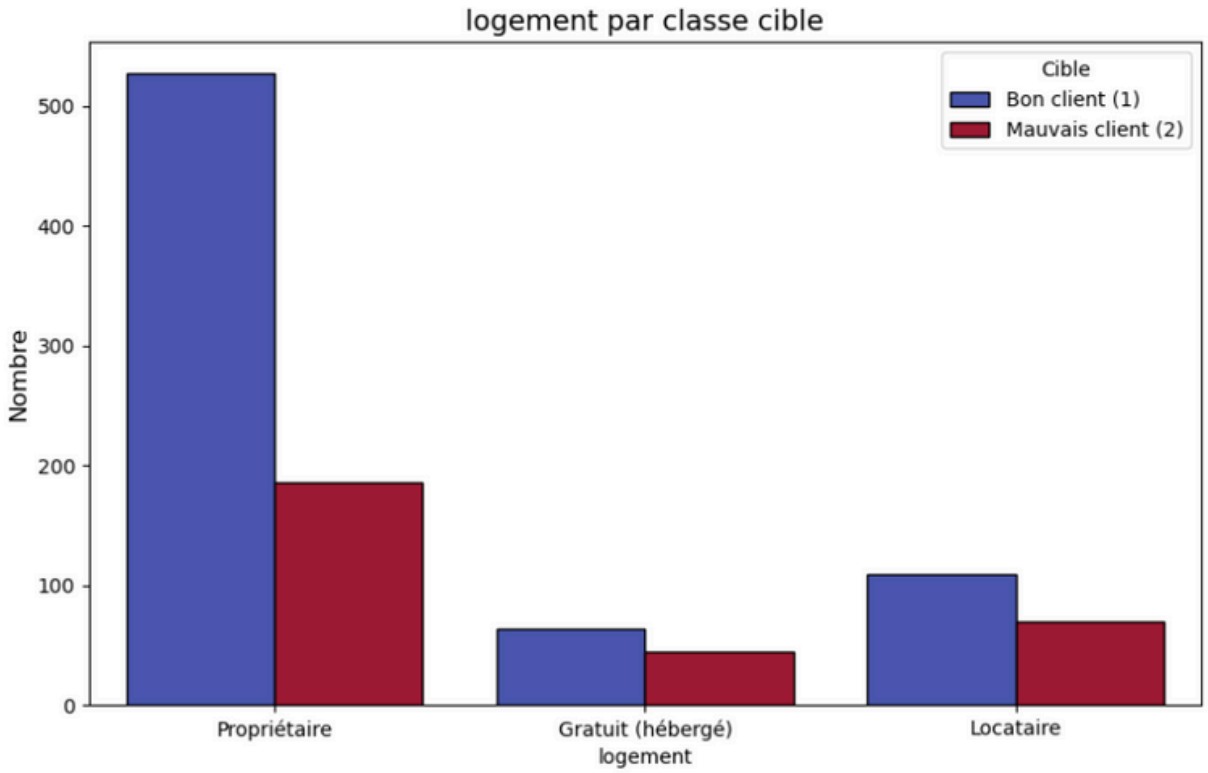
Dans le graphique montrant la relation entre montant et durée_mois, les mauvais clients (points rouges) ont tendance à se concentrer dans des zones de montants plus élevés et de durées plus longues, tandis que les bons clients (points bleus) sont principalement situés autour des montants faibles et des durées courtes. Cela suggère une corrélation entre des crédits plus risqués et des montants plus importants ou des durées prolongées. Pour la relation entre âge et montant, on remarque que les clients plus jeunes, quelle que soit leur catégorie, empruntent souvent des montants moins élevés, tandis que les clients plus âgés montrent une plus grande variabilité.

En ce qui concerne les graphiques de densité sur la diagonale, ils permettent de visualiser la répartition de chaque variable pour les bons et mauvais clients. Par exemple, pour la variable montant, les bons clients ont une densité plus forte autour des montants faibles (moins de 5000), tandis que les mauvais clients ont une répartition plus étalée, avec une densité non négligeable pour les montants élevés. Cela reflète des différences significatives dans le comportement des deux groupes.

En conclusion, ce graphique montre que les mauvais clients ont tendance à emprunter des montants plus élevés et sur des durées plus longues, tandis que les bons clients se concentrent davantage sur des montants modestes et des durées courtes. Les graphiques de densité mettent en évidence ces différences de manière globale et facilitent l'interprétation des distributions.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



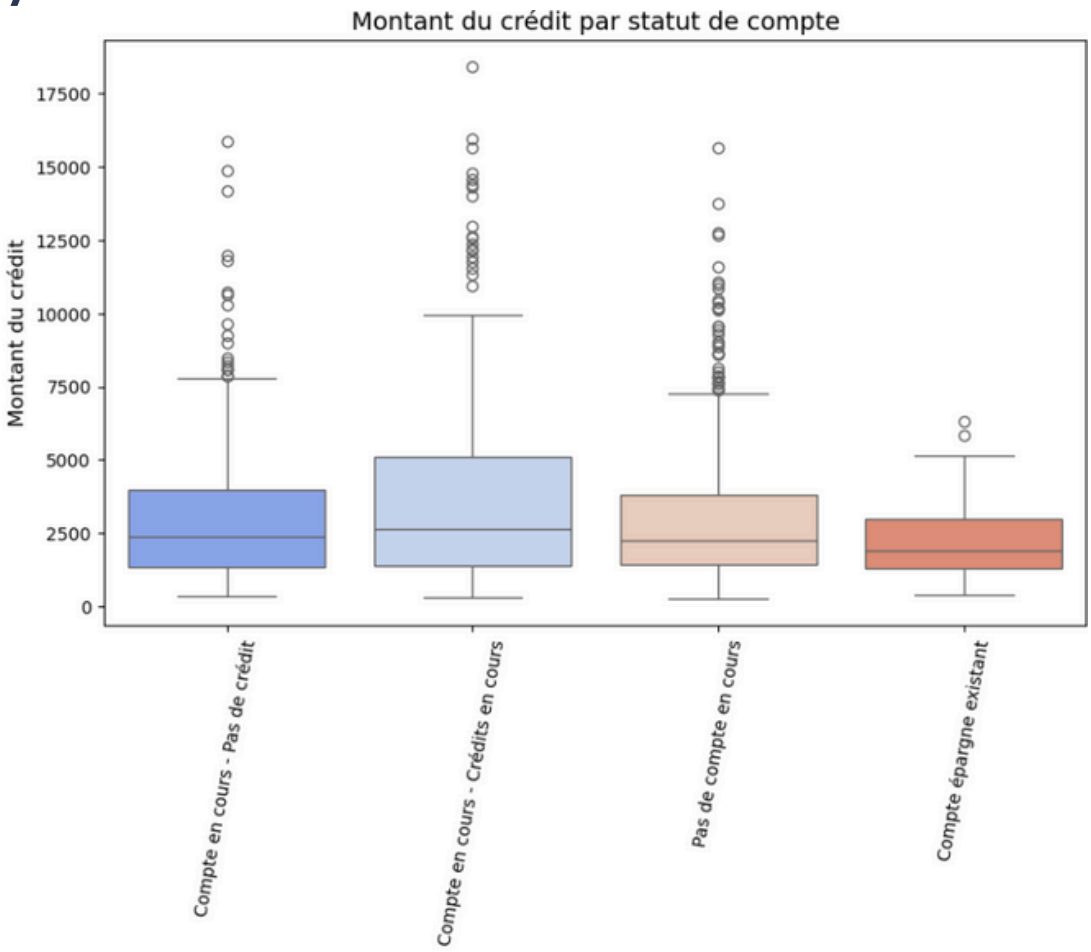
Ce graphique montre la répartition des clients en fonction de leur type de logement (propriétaire, locataire ou hébergé gratuitement) et de leur classification en bons ou mauvais clients. Il est utile pour comprendre si le type de logement a un lien avec le comportement des clients, ce qui peut aider à évaluer le risque financier lié à ces profils.

Pour les bons clients (Cible 1), la majorité sont propriétaires, avec plus de 500 individus dans cette catégorie. Les locataires sont bien moins nombreux, suivis des clients hébergés gratuitement. Pour les mauvais clients (Cible 2), les proportions restent similaires : les propriétaires restent majoritaires avec environ 200 individus, mais les locataires et les clients hébergés gratuitement sont proportionnellement plus nombreux dans ce groupe.

En conclusion, ce graphique montre que les propriétaires sont généralement plus nombreux parmi les bons clients, tandis que les locataires et les personnes hébergées gratuitement sont davantage représentés proportionnellement parmi les mauvais clients. Cela pourrait indiquer que le type de logement est un indicateur de stabilité financière et de fiabilité.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



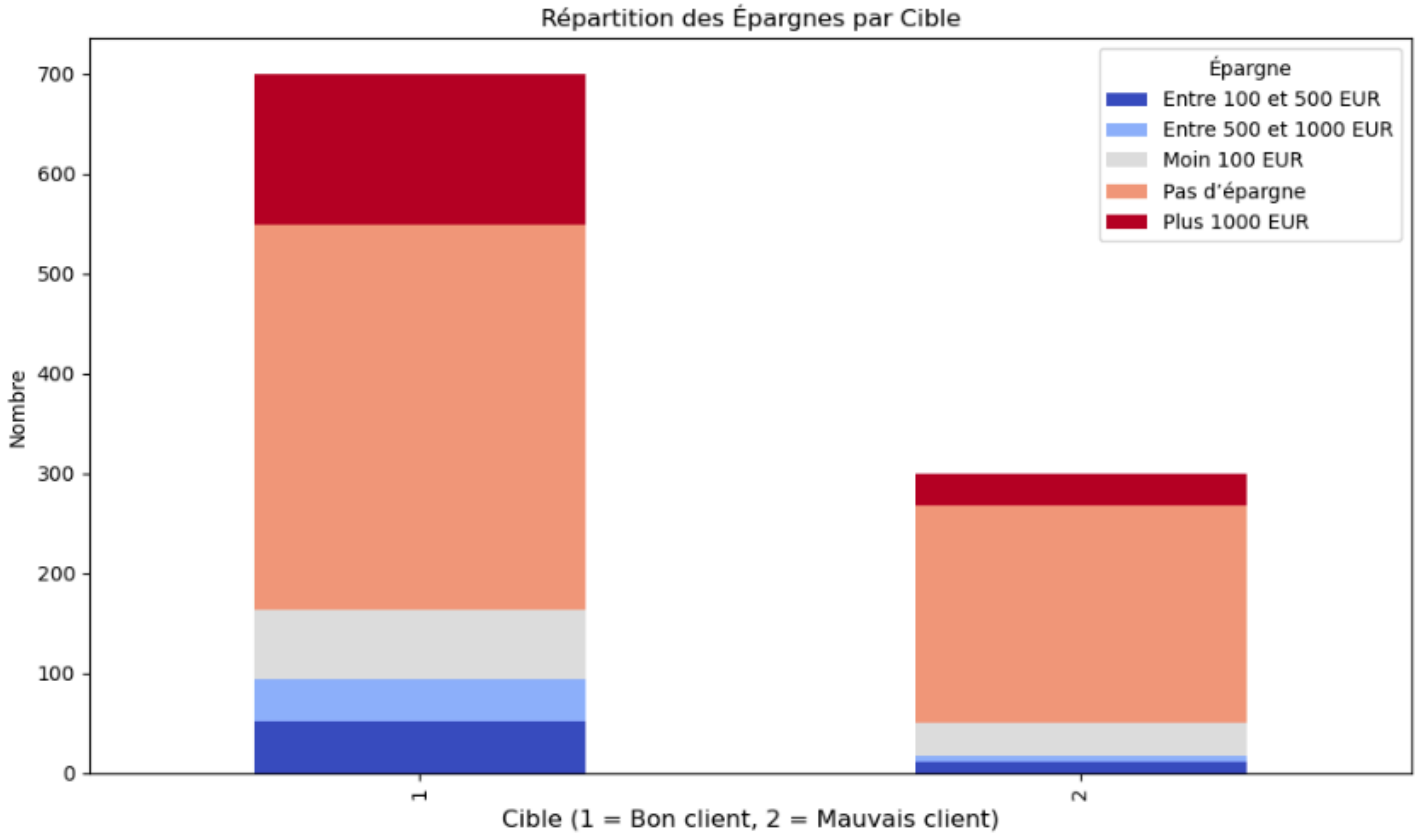
Ce graphique, sous forme de boîtes à moustaches, montre la répartition des montants de crédit selon le statut de compte des clients. Il est utile pour analyser si le statut de compte influence le montant emprunté, ce qui peut aider à comprendre les comportements financiers des clients en fonction de leur situation bancaire.

Les clients ayant un compte en cours avec crédits en cours présentent une dispersion plus importante des montants empruntés, avec une médiane plus élevée que les autres groupes et de nombreux prêts dépassant 10 000 unités. Les clients avec un compte en cours mais pas de crédit ont également une médiane relativement élevée, bien que les montants soient légèrement moins dispersés. À l'inverse, les clients sans compte en cours ou avec un compte d'épargne existant empruntent des montants plus faibles, avec des médianes plus basses.

En conclusion, ce graphique révèle que les clients avec des crédits en cours ont tendance à emprunter des montants plus élevés, ce qui peut indiquer un profil plus risqué ou une capacité d'emprunt plus importante. Ces informations sont utiles pour évaluer les comportements des clients en fonction de leur statut bancaire.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée



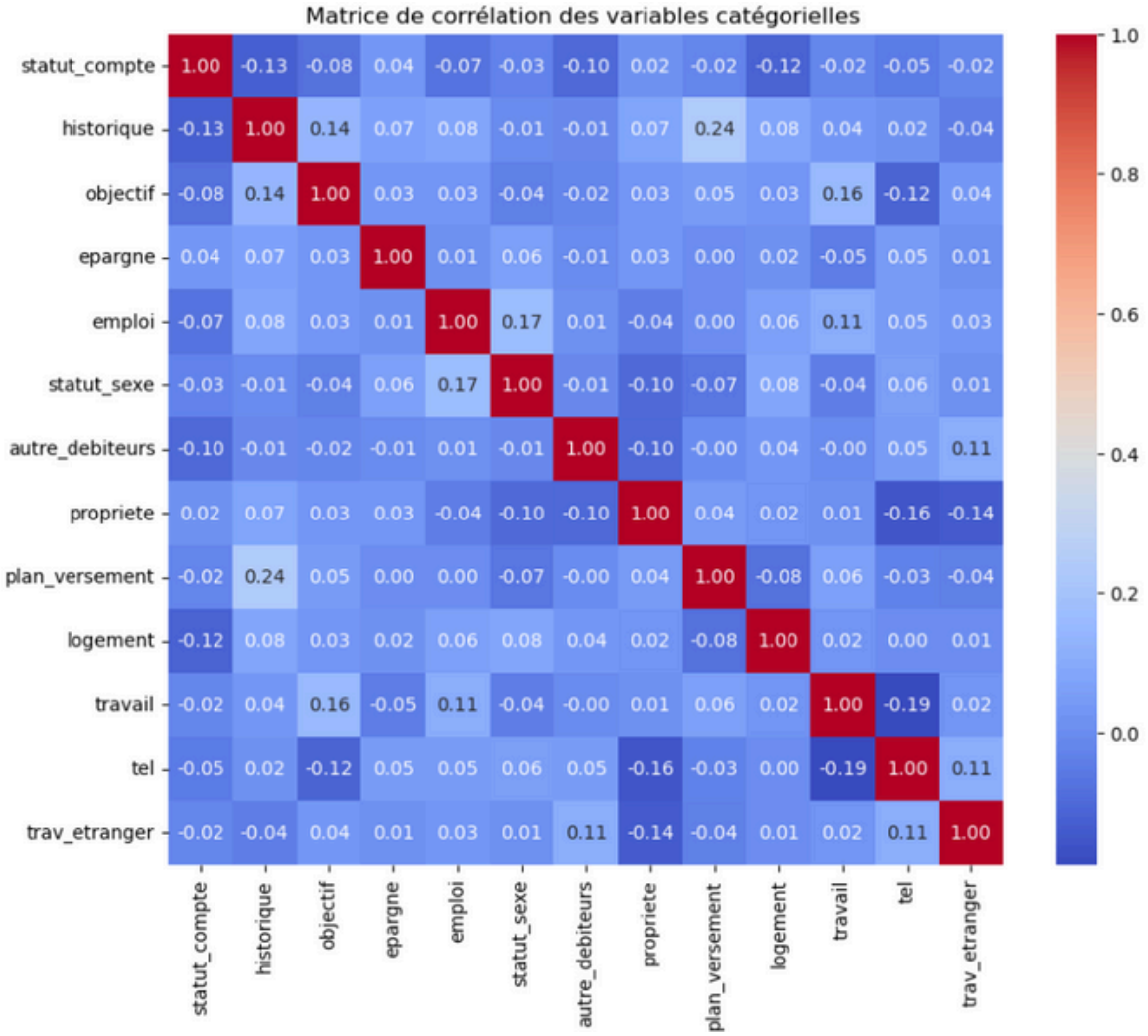
Ce graphique montre la répartition des niveaux d'épargne des clients selon leur classification en bons (Cible 1) ou mauvais clients (Cible 2). Il est utile pour identifier des tendances financières entre les deux groupes et comprendre si le niveau d'épargne joue un rôle dans la classification des clients, ce qui peut être crucial pour évaluer leur stabilité financière.

Pour les bons clients (Cible 1), une majorité représentée en orange, n'a pas d'épargne. Cependant, une proportion significative dispose de plus de 1000 EUR (en rouge), ce qui peut refléter une certaine capacité financière chez certains bons clients. Les autres niveaux d'épargne, comme ceux compris entre 100 et 500 EUR (en bleu clair), sont minoritaires. Concernant les mauvais clients (Cible 2), on observe une tendance similaire : la majorité n'a pas d'épargne, mais la proportion de ceux ayant plus de 1000 EUR est plus faible par rapport aux bons clients.

En conclusion, ce graphique révèle que l'absence d'épargne est fréquente dans les deux groupes, mais les bons clients montrent une proportion légèrement plus élevée d'épargnes importantes, ce qui peut indiquer un meilleur potentiel financier. Ces informations peuvent guider les décisions en matière d'attribution de crédit.

ANALYSE EXPLORATOIRE DES DONNÉES

II/Analyse Bivariée

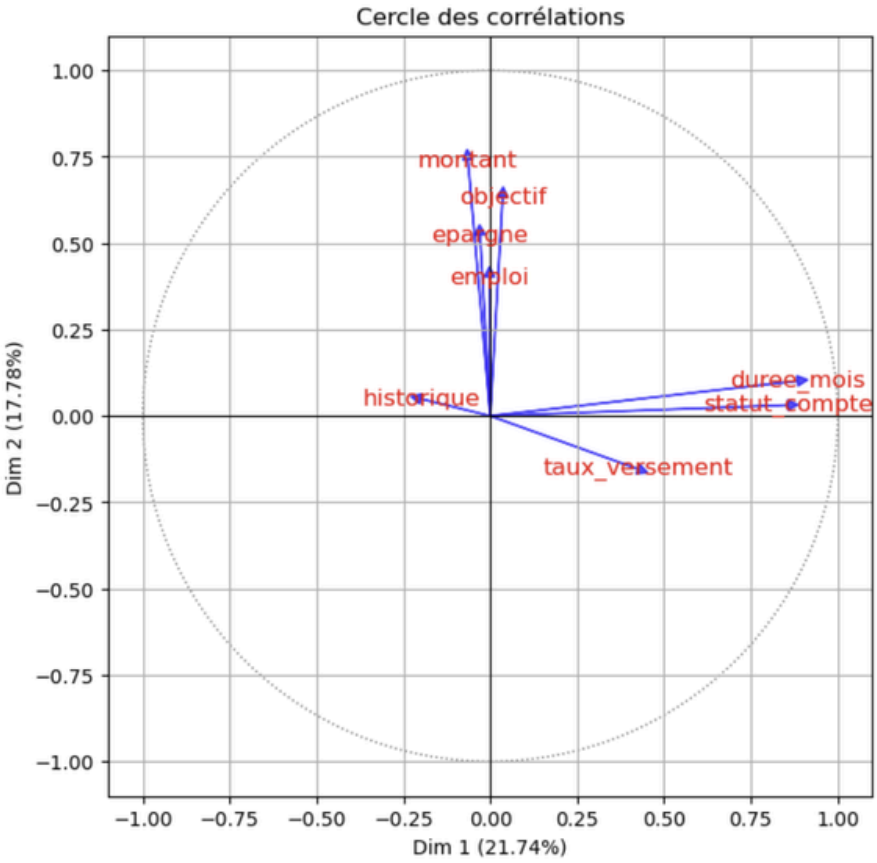


Cette matrice de corrélation met en lumière les relations entre les variables catégorielles. Chaque valeur indique le degré de corrélation entre deux variables, allant de -1 (corrélation négative parfaite) à 1 (corrélation positive parfaite). Ce graphique est utile pour identifier les relations significatives ou indépendantes entre les variables, ce qui peut aider à simplifier les modèles ou à comprendre les facteurs influents.

Par exemple, on observe une corrélation de 0,24 entre les variables historique et plan_versement, ce qui suggère que les modalités d'historique de crédit sont légèrement liées au choix du plan de versement. De même, la variable statut_sexe montre une corrélation modérée de 0,17 avec la variable emploi, indiquant une association entre le type d'emploi et le sexe des clients. En revanche, la plupart des autres variables, comme propriété et travail, présentent des corrélations faibles (proches de 0), ce qui indique une indépendance.

En conclusion, cette matrice révèle que certaines relations, bien que faibles, pourraient être explorées davantage, comme entre l'historique et le plan de versement. Elle permet également de confirmer que plusieurs variables sont faiblement corrélées, ce qui indique qu'elles capturent des dimensions différentes des données.

L'ANALYSE EN COMPOSANTES PRINCIPALES



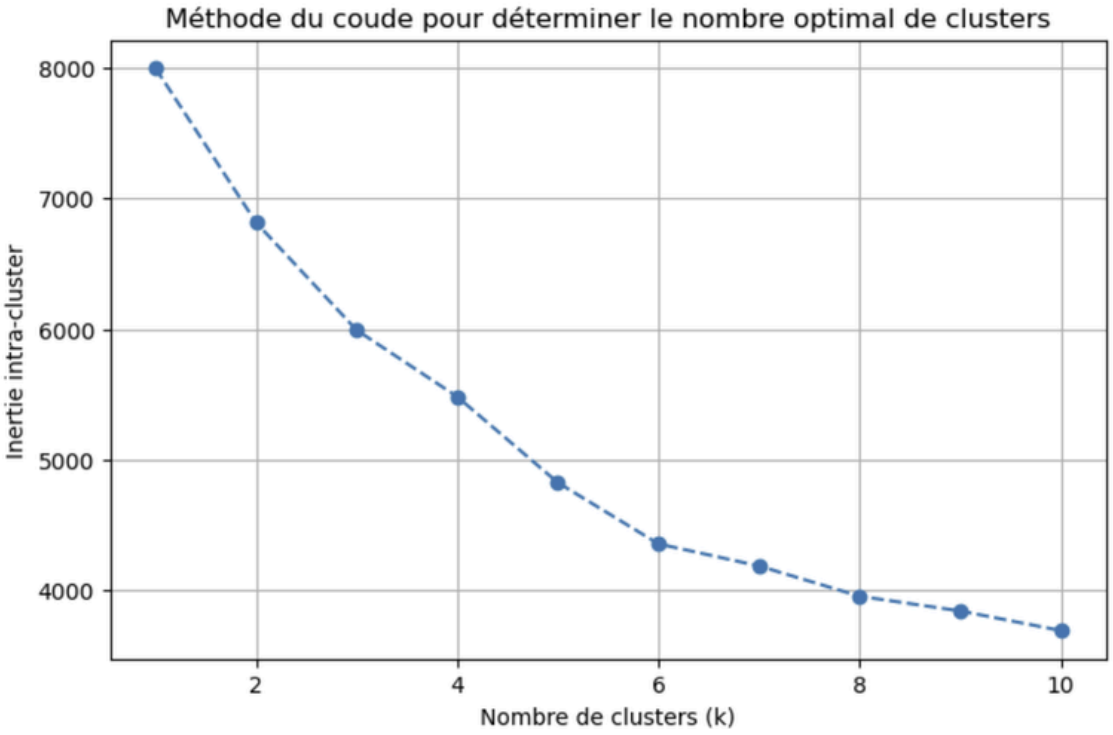
L'ACP est une méthode statistique qui permet de réduire la dimensionnalité des données tout en conservant un maximum d'informations. Elle aide à identifier les relations entre les variables et à regrouper celles qui sont corrélées. Cela simplifie l'interprétation et permet de visualiser les données dans un espace réduit, ici les deux premières dimensions. Ces dimensions expliquent respectivement 21,74 % et 17,78 % de la variance totale.

Le graphique montre la projection des variables initiales dans un plan formé par les deux premières dimensions principales. Les flèches représentent la corrélation entre les variables et les composantes principales. Plus une variable est éloignée de l'origine, plus elle est bien représentée dans ce plan. Les variables proches l'une de l'autre sont corrélées positivement, tandis que celles qui sont opposées indiquent une corrélation négative.

Les variables associées à la première dimension (Dim 1) comme durée_mois, statut_compte et taux_versement semblent refléter une dynamique commune probablement liée aux aspects financiers des crédits. La seconde dimension (Dim 2) regroupe des variables comme montant, objectif, épargne et emploi, qui peuvent être liées au profil socio-économique des individus. Ce graphique révèle aussi que certaines variables sont redondantes ou fortement corrélées, ce qui peut justifier une réduction du nombre de variables pour simplifier l'analyse.

L'ACP permet ici de détecter des groupes de variables et de mieux comprendre leur structuration autour des deux dimensions principales. Cela prépare également les données pour une étape de clustering comme k-means, en exploitant les composantes principales pour regrouper les individus selon leurs similarités. Ce graphique est essentiel pour guider la réorganisation des données et pour simplifier les prochaines étapes d'analyse.

CLUSTERING AVEC K-MEANS

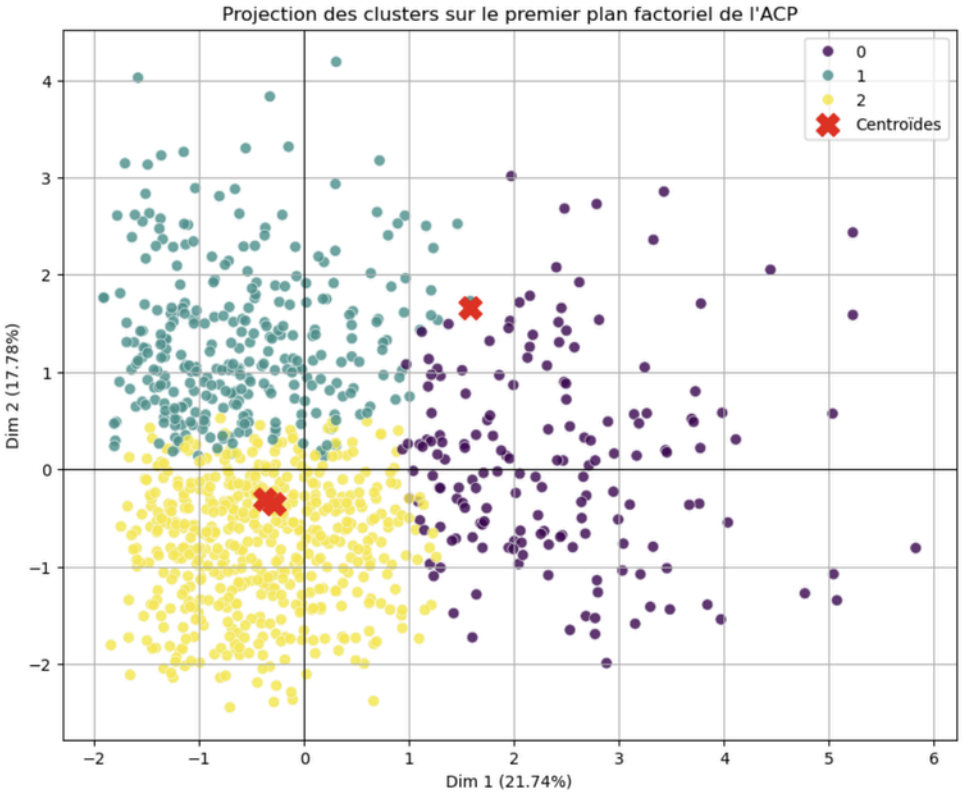


Le graphique présente l'application de la méthode du coude, une technique utilisée pour déterminer le nombre optimal de clusters dans le cadre d'une analyse de classification. Cette méthode repose sur l'évaluation de l'inertie intra-cluster, qui mesure la somme des distances au carré des points au centre de leur cluster. L'objectif est d'identifier un équilibre entre le nombre de clusters et la qualité de la segmentation.

Sur le graphique, on observe une diminution rapide de l'inertie entre les valeurs de k=1k = 1k=1 et k=3k = 3k=3, suivie d'une réduction plus modérée au-delà de k=3k = 3k=3. Cela traduit un rendement décroissant lorsqu'on ajoute davantage de clusters. Ce phénomène est caractéristique de ce qu'on appelle le « coude », qui se situe ici autour de k=3k = 3k=3. Ce point marque une transition où l'ajout de nouveaux clusters n'améliore plus significativement la qualité de la classification.

Le choix de k=3k = 3k=3 comme nombre optimal de clusters se justifie par cet équilibre. À ce niveau, on obtient une segmentation à la fois simple et efficace. Ajouter des clusters supplémentaires augmenterait la complexité de l'analyse sans apporter de gains notables en termes de cohérence ou de compacité des groupes. Ce choix est particulièrement pertinent dans un contexte où une interprétation claire des clusters est essentielle pour orienter les décisions futures. La segmentation en trois groupes offre ainsi une structure significative et exploitable pour des analyses ou actions subséquentes.

CLUSTERING ET ACP



Ce graphique illustre la projection des clusters identifiés sur le premier plan factoriel d'une Analyse en Composantes Principales (ACP). Les points représentent les observations projetées selon les deux premières dimensions principales, tandis que les couleurs indiquent les trois clusters obtenus grâce à la segmentation. Les croix rouges marquent les centroïdes, qui correspondent aux centres de gravité de chaque cluster.

L'ACP a permis de réduire la dimensionnalité des données initiales tout en conservant une part importante de la variance, comme en témoigne le pourcentage de variance expliqué par les deux axes : 21,74 % pour la première dimension (Dim 1) et 17,78 % pour la seconde (Dim 2). Ces deux axes représentent donc les directions les plus explicatives de la dispersion des données.

Ce graphique met en évidence la structure des clusters dans cet espace réduit. Les trois groupes sont clairement séparés, ce qui confirme la pertinence du choix de k=3k = 3k=3 effectué précédemment. Chaque cluster montre une certaine homogénéité, car les points d'un même groupe sont regroupés autour de leur centroïde.

La répartition des clusters suggère des différences significatives entre les groupes selon les dimensions principales, ce qui peut être interprété comme des caractéristiques ou comportements distincts des observations. Ce type de visualisation facilite la compréhension des résultats et aide à justifier l'analyse en mettant en lumière les spécificités des clusters. Ces informations pourront être exploitées pour des analyses complémentaires ou des prises de décision stratégiques basées sur la segmentation obtenue.

CONCLUSION

PARTI 1: La partie I de ce rapport a permis d'explorer et d'analyser les données des clients bancaires afin d'identifier les principales tendances et relations entre les variables. Les analyses univariées ont révélé des éléments significatifs, tels que l'importance de l'épargne et la stabilité financière chez les bons clients, ainsi que l'influence du statut de compte sur leur classification.

Les analyses bivariées ont enrichi cette compréhension en mettant en évidence des corrélations intéressantes, comme le lien entre l'historique de crédit et les plans de versement, ou encore l'impact des montants empruntés sur le comportement des clients. Ces relations soulignent des indicateurs clés dans la gestion du risque client.

Enfin, l'examen de la matrice de corrélation a permis de détecter des relations significatives, bien que globalement faibles, entre certaines variables catégorielles et continues. Ces résultats montrent que certaines caractéristiques, comme le montant emprunté, le statut de compte ou les plans de versement, jouent un rôle crucial dans l'évaluation du risque client.

Ces analyses posent les bases d'une segmentation plus approfondie et mettent en lumière la nécessité d'intégrer des modèles analytiques plus complexes pour mieux comprendre les comportements des clients. Elles constituent une fondation solide pour orienter les stratégies des institutions financières en matière de gestion des risques et d'optimisation des décisions commerciales.

PARTI 2: La partie II de ce rapport a permis de mettre en œuvre une approche de segmentation des clients basée sur l'analyse des données et des algorithmes de classification non supervisée. En appliquant la méthode du coude, nous avons déterminé que k=3k = 3k=3 était le nombre optimal de clusters, garantissant un bon compromis entre la complexité du modèle et la qualité de la segmentation.

La projection des clusters sur le plan factoriel de l'Analyse en Composantes Principales (ACP) a confirmé la validité de ce choix. Les trois clusters identifiés présentent des groupes homogènes et bien distincts, reflétant des différences significatives dans les caractéristiques comportementales et financières des clients. Cette visualisation a facilité l'interprétation des résultats en mettant en évidence des profils distincts qui peuvent être exploités pour des décisions stratégiques ciblées.

Ces résultats offrent une segmentation exploitable qui peut guider les institutions financières dans la personnalisation de leurs stratégies, qu'il s'agisse de minimiser les risques ou d'optimiser la relation client. La méthodologie utilisée dans cette partie illustre l'importance des outils analytiques pour transformer des données brutes en informations directement applicables aux objectifs stratégiques de gestion des risques et de fidélisation des clients.