Rapport de projet : programmation statistique avec Python

ALI KINAZA
AOUIMEUR ASSIA
ZIDAN NADIA

Sommaire

ANALYSE DESCRIPTIVE

1- Description du jeu de données (p.4)

2- Analyse univarié des variables (p.7)

3- Analyse bivarié des variables (p.9)

4- Analyse multivarié de variables (p.13)

Prologue

Dans le cadre d'un projet d'analyse exploratoire sur des données d'assurance, une étude est réalisée sur des données réelles. L'objectif est d'analyser un jeu de données concernant des crédits allemands, qui décrit les détails financiers et bancaires des clients d'une banque.

L'ensemble de l'analyse est effectué avec Python, et la base de données est directement accessible via la bibliothèque **ucimlrepo**.
L'objectif principal de l'analyse est de déterminer, grâce à de nombreux facteurs caractérisant les individus, quels sont les clients potentiellement à risque ou non. Une variable cible est présente dans la base de données pour cet objectif :

- Elle contient la valeur 1 si le client est considéré comme "bon".
- Elle contient la valeur 2 si le client est considéré comme "à risque".

L'analyse vise donc à identifier les facteurs permettant de prédire si un client est bon ou non, afin de minimiser les risques de non-remboursement des crédits pour les banques.

Tout d'abord, une description des données est réalisée afin de comprendre les variables disponibles et leur lien avec les individus (variables utilisées, variables supplémentaires, etc.).
Ensuite, une analyse en composantes principales (ACP) est effectuée, suivie par des techniques de clustering telles que le K-means et la classification hiérarchique ascendante (CAH), avant de conclure.

Partie 1 : Manipulation, analyse et visualisation des données statlog_german_credit_data

Dans cette première partie, nous allons manipuler, analyser et visualiser les données issues du jeu de données Statlog German Credit Data, qui décrit les détails financiers et bancaires de clients d'une banque. Ces données sont accessibles depuis la base UCI Machine Learning Repository et ont été récupérées en utilisant la bibliothèque Python ucimlrepo.

L'objectif est de comprendre la structure des données, d'explorer leurs caractéristiques et de préparer des visualisations pertinentes pour faciliter l'analyse. Nous avons commencé par importer le jeu de données, renommer les colonnes pour une meilleure lisibilité, et effectuer une description statistique préliminaire.

L'ensemble de données se compose de 1 000 individus et 20 variables, dont :

- 7 variables numériques (de type entier),
- 13 variables catégorielles.

La dernière colonne correspond à la variable cible :

- 1 : bon client (faible risque de crédit),
- 2 : mauvais client (risque élevé de crédit).

Il est important de noter que cette base de données ne contient aucune valeur manquante.

Le tableau descriptif des variables est présenté ci-dessous, permettant d'avoir une vue d'ensemble des caractéristiques du jeu de données.

Structure des données pour l'analyse :

- X (features) : Ces variables descriptives sont utilisées comme entrée pour entraîner les modèles de machine learning. Elles contiennent les informations nécessaires pour expliquer ou influencer la variable cible.
- y (target) : Cette variable cible représente l'indicateur principal de l'analyse. Elle détermine si un client est un "bon" ou un "mauvais" risque de crédit.

En résumé, l'objectif est de prédire y (la cible) à partir de X (les descripteurs). Toutes les variables explicatives servent d'indicateurs pour évaluer si un individu est à risque ou non, dans le cadre d'un modèle de prédiction.

Nom de la variable qualitative	Description
Statut_compte	Statut du compte (ex : A11, A12) indiquant l'historique bancaire
Tel	Présence d'un téléphone (oui/non).
Trav_etranger	Statut lié au travail à l'étranger.
Historique	Historique de crédit (ex : A32, A34).
Logement	Type de logement (ex : propriétaire, locataire).
Plan_versement	Plans de versement proposés.
Propriete	Type de bien possédé (ex : immobilier, voiture).
Autre_debiteurs	Autres débiteurs ou garants pour le crédit.
Statut_sexe	Statut personnel et genre (ex : homme célibataire, femme mariée).
Emploi	Durée d'emploi (ex : stable, moins d'1 an).
Epargne	Épargne ou actifs disponibles.
Travail	Type d'emploi occupé
Objectif	Le but recherché avec ce crédit (voiture, logement...)
Nom de la variable quantitative	**Description**
Duree_mois	Durée de remboursement du crédit en mois.
Montant	Montant du crédit demandé.
Taux_versement	Taux de remboursement par rapport au revenu.
Residence	Durée de résidence actuelle.
Âge	Âge du client
N_credit	Nombre de crédits existants pour le client.
N_p_charge	Nombre de personnes à charge.

Classification des variables : qualitatives et quantitatives

Pour une analyse cohérente et adaptée, nous avons décidé de scinder les variables du jeu de données en deux catégories :
Variables qualitatives, représentant des informations catégoriques ou textuelles (exemple : statut_sexe, historique ,etc...).
Variables quantitatives, représentant des valeurs numériques continues ou discrètes (exemple : montant, age, etc...).
Cette classification permet de choisir les outils d'analyse et de visualisation appropriés, d'assurer une préparation adéquate des données pour les modèles, et d'apporter une meilleure clarté à l'interprétation des résultats.

```
Variables qualitatives : ['statut_compte', 'historique', 'objectif', 'epargne', 'emploi', 'statut_sexe', 'autre_debiteurs', 'propriete', 'plan_versement', 'logement', 'travail', 'tel', 'trav_etranger']
Variables quantitatives : ['duree_mois', 'montant', 'taux_versement', 'residence', 'age', 'n_credit', 'n_p_charge', 'Target']
```

Ce processus nous ait utile tout le long de notre démarche notamment pour les analyses graphiques univariées et bi-variées de chaque variable et également lorsque nous feront l'algorhitme de l'ACP (analyse en composantes principales).

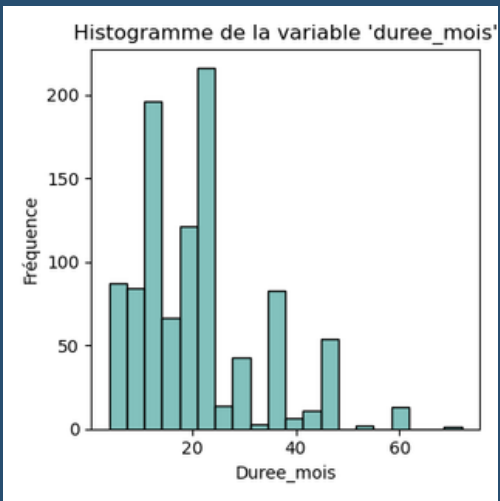
Partie 2 : Analyse univariée

Dans un premier temps, afin de comprendre les 21 variables à analyser, une étude globale univariée a été réalisée. Tout d'abord, avec l'instruction X.describe(include="all") sur Python, qui effectue une analyse préliminaire des données. Les indicateurs statistiques basiques, tels que la médiane ou encore le nombre de modalités par variable selon la nature des variables (quantitatives ou qualitatives), sont ainsi présents afin de se donner une première idée des tendances (cf. code Jupyter).

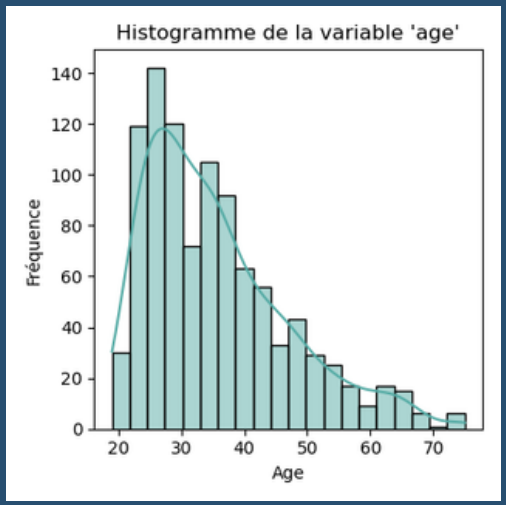
Suite à cela, une visualisation univariée de chaque variable a été réalisée à partir d'une boucle sur chaque variable, en prenant en compte la nature de la variable, afin de pouvoir sélectionner les variables qui semblent pertinentes pour réaliser les analyses bivariées.

Voici les visualisations qui ont retenu notre attention (cf. code Jupyter pour voir l'ensemble des résultats).

Visualisation univariée

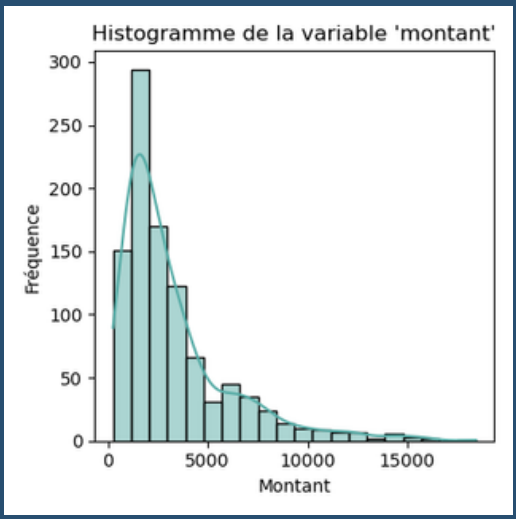


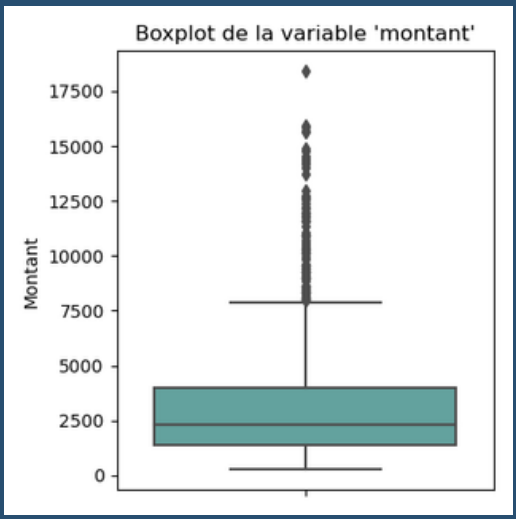
Cet histogramme représente la distribution des fréquences des produits en fonction de leur durée. On constate que les crédits ont tendance à durer entre 10 et 25 mois.



L'histogramme montre que la majorité des individus ont entre 30 et 40 ans, avec une distribution décroissante au-delà de cet âge. Les jeunes adultes (20-30 ans) sont également bien représentés, tandis que les plus de 60 ans sont rares.

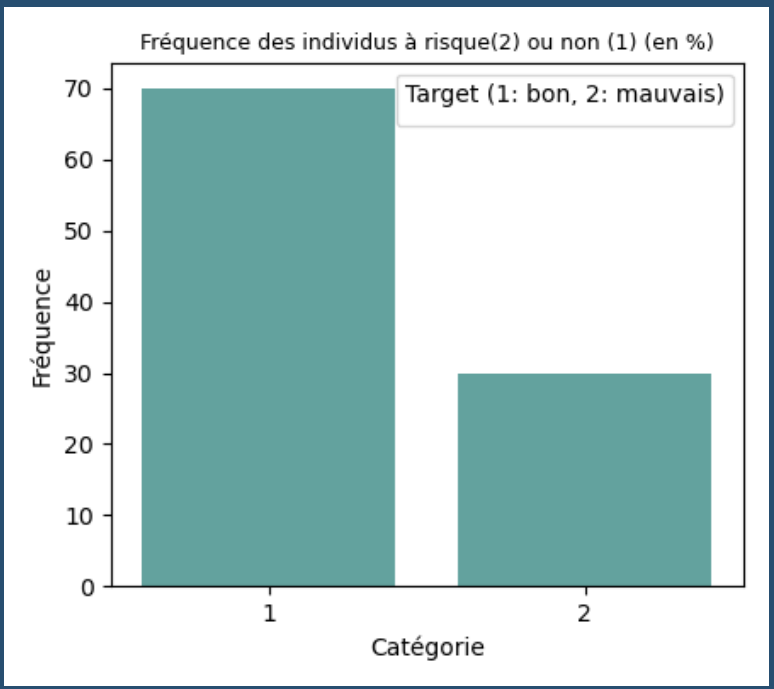
Analyse univariée visualisation



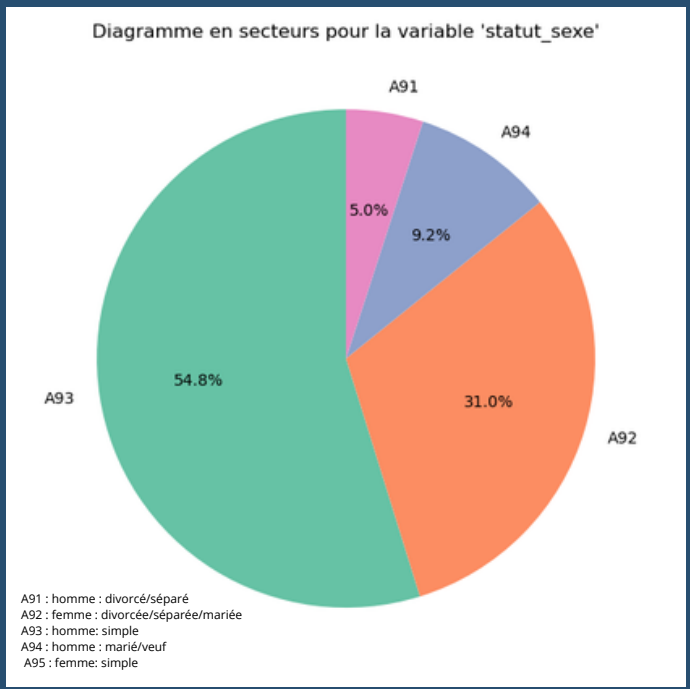


Ici, la distribution ainsi que la dispersion du montant des crédits accordés sont présentées. On constate que la majorité des crédits accordés sont inférieurs à 5000 euros.

Variable qualitatives



La variable d'intérêt est ici analysée afin de déterminer la proportion d'individus dans la dataframe considérés comme à risque ou non.



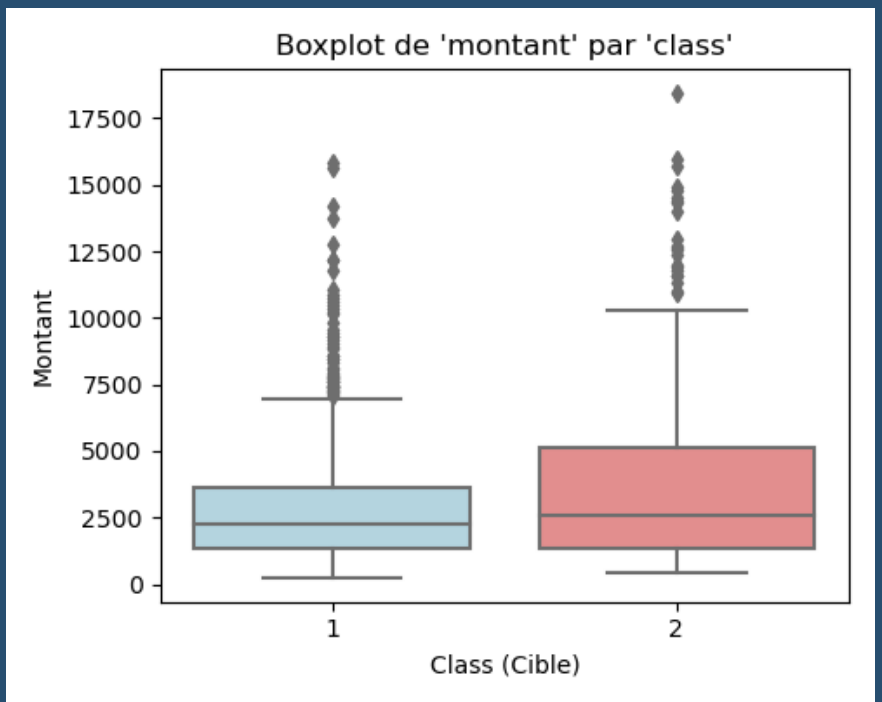
Répartition de statut_sexe. La majorité (54,8 %) sont des hommes célibataires (A93), suivis par 31 % de femmes mariées/divorcées (A92). Les hommes mariés/veufs (A94) et divorcés/séparés (A91) représentent respectivement 9,2 % et 5 %.

Partie 3 : Visualisation Analyse bivariée

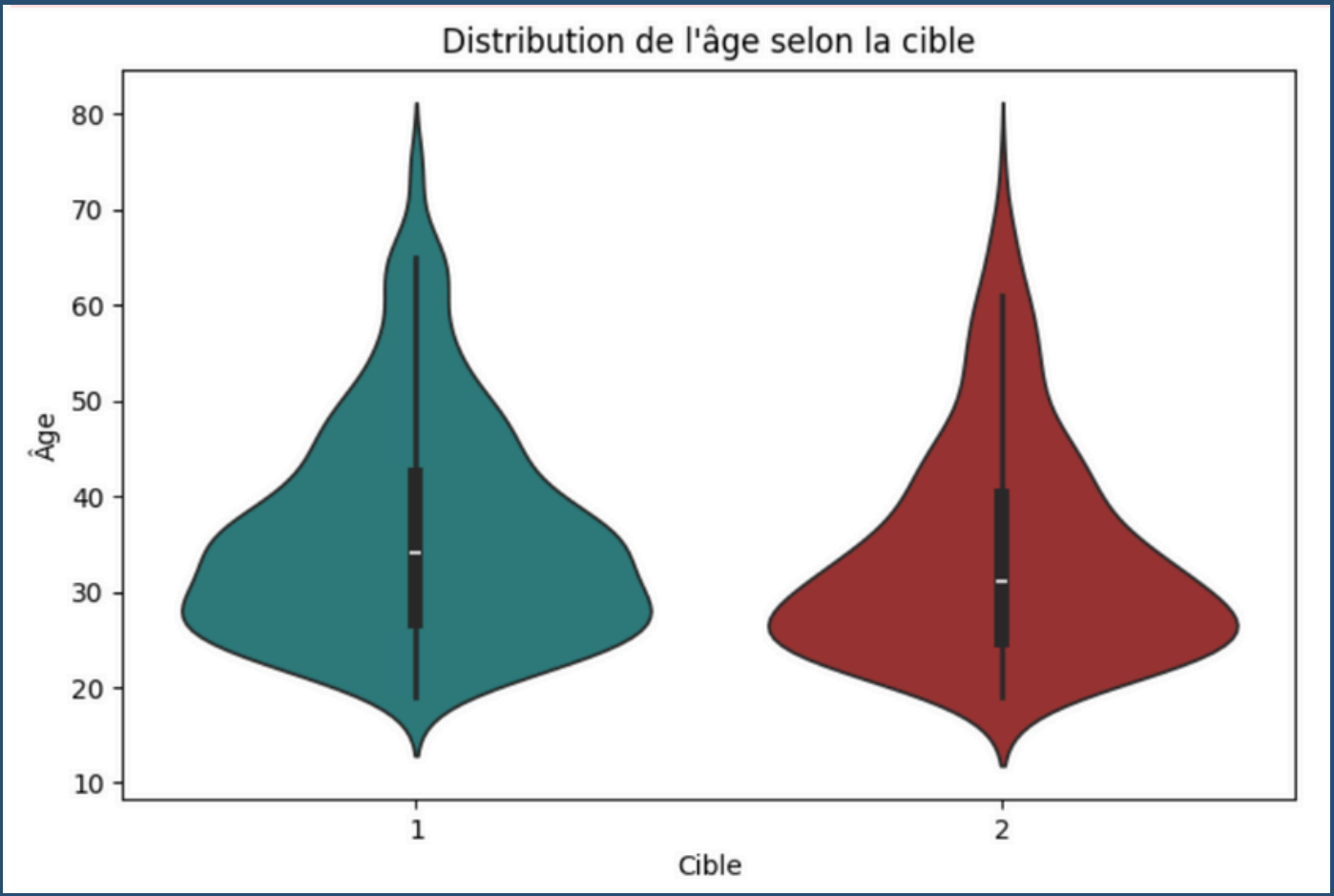
Afin de pousser notre analyse, une visualisation avec les variables croisées est faite. L'ensemble des graphiques fait sont accessible dans le code Jupiter mais ici nous avons retenue les graphiques qui nous semblent intéressants à commenter.

La cible 2 présente une durée médiane plus élevée et une plus grande dispersion, tandis que la cible 1 regroupe des crédits majoritairement concentrés sur des durées courtes.

Les montants médians sont comparables entre les deux cibles, mais la cible 2 montre une distribution plus étendue avec davantage d'outliers pour des crédits élevés.

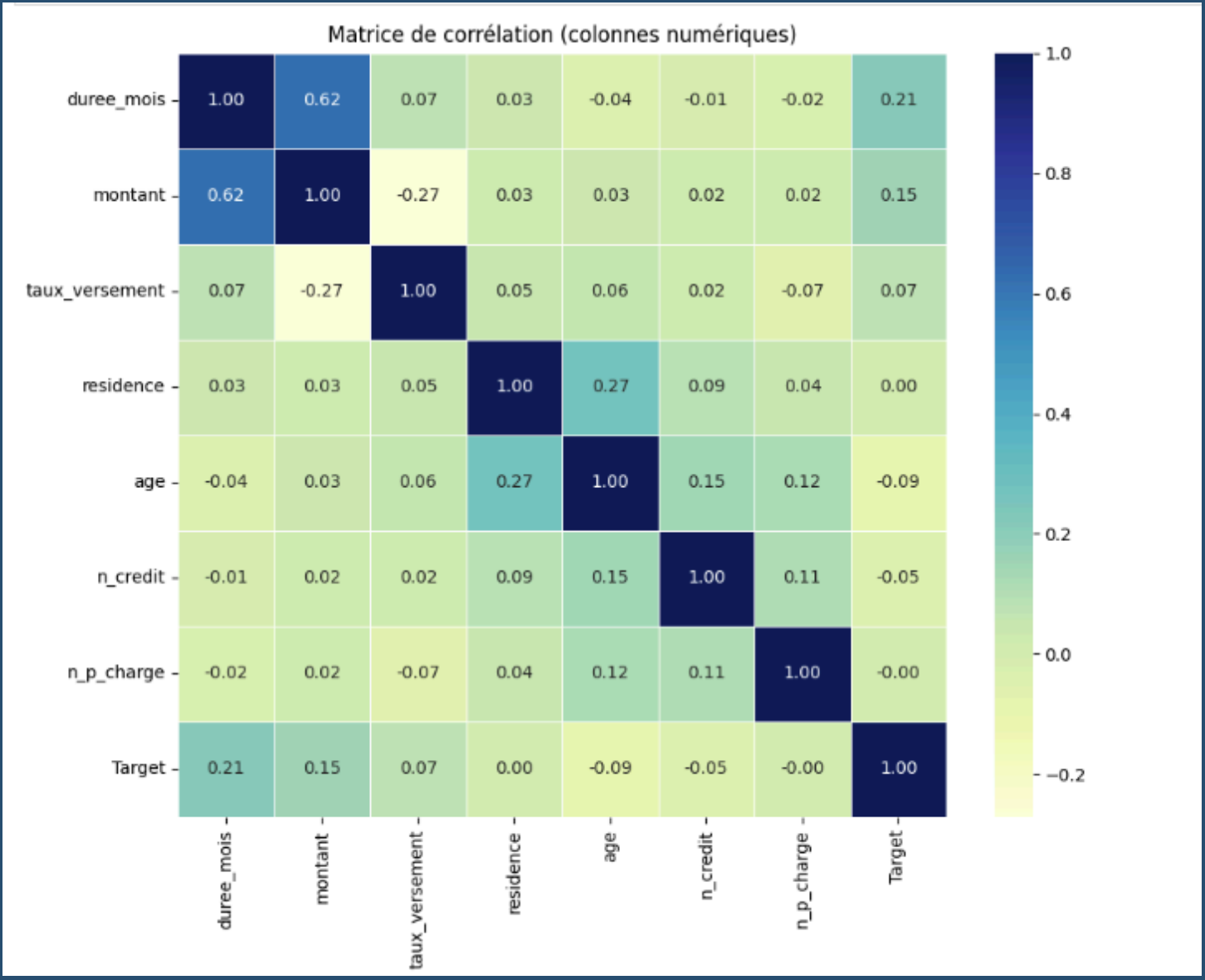


Visualisation Analyse bivariée



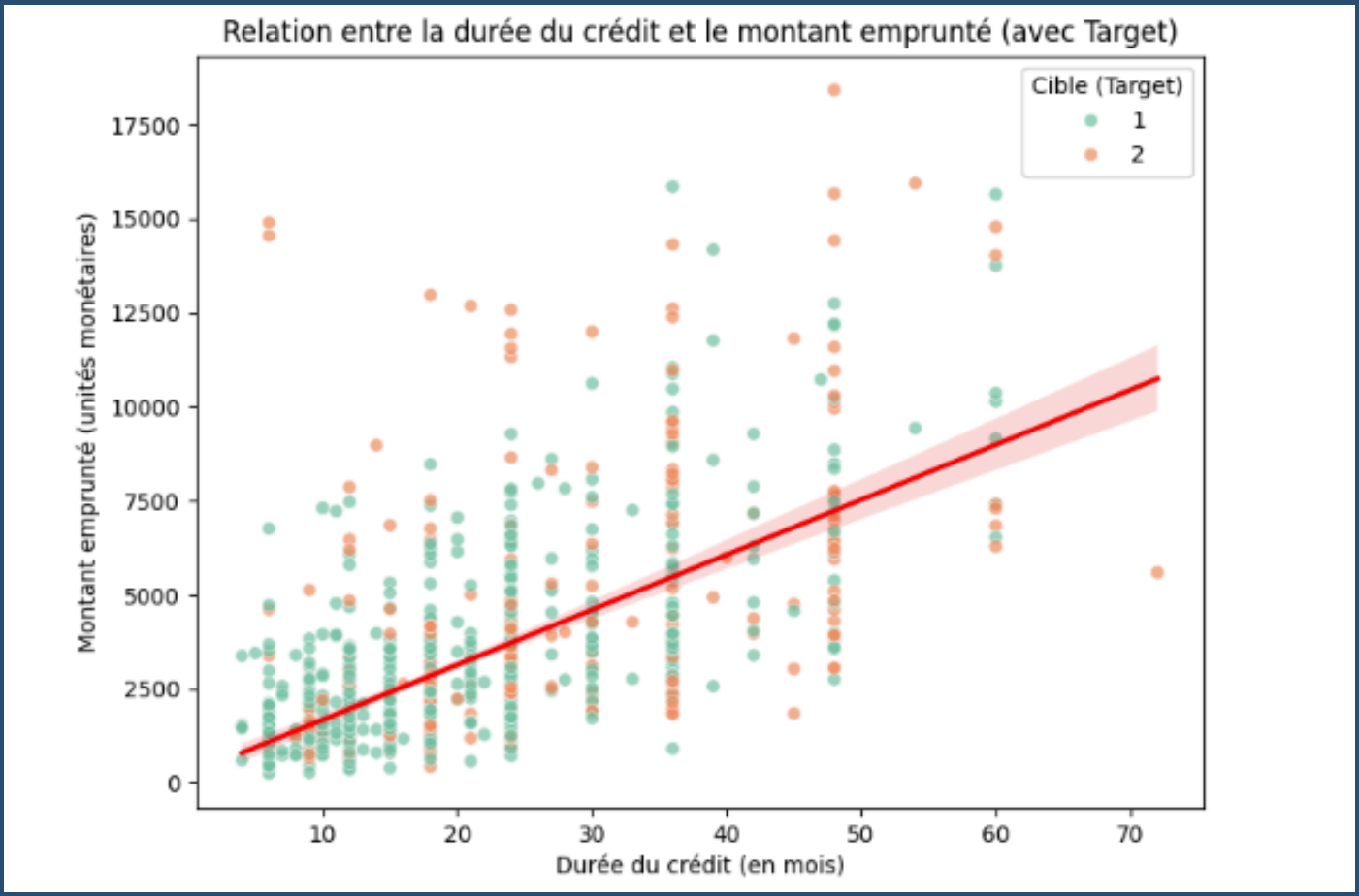
Ce graphique en violon montre que la distribution de l'âge est similaire pour les deux cibles, avec une médiane située entre 35 et 40 ans. Toutefois, la **cible 1** présente une dispersion plus importante, regroupant davantage de jeunes (20-30 ans) et de personnes âgées (70-80 ans). En revanche, la **cible 2** est plus symétrique, avec une concentration autour de la médiane. Ces observations indiquent qu'il n'y a pas de lien évident entre l'âge et la cible, bien que les extrêmes soient plus fréquents dans la **cible 1.**

Visualisation Analyse bivariée



La matrice de corrélation montre que la durée du crédit (duree_mois) et le montant emprunté (montant) ont une corrélation positive (0.62), indiquant que les crédits de longue durée sont souvent associés à des montants plus élevés. En ce qui concerne la cible (Target), les corrélations avec les autres variables sont faibles, avec un lien léger avec la durée du crédit (0.21) et le montant (0.15). Ces résultats suggèrent que les variables présentes dans l'analyse n'expliquent que partiellement la cible, et d'autres facteurs pourraient influencer davantage le comportement des emprunteurs. Pour finir, on note une corrélation négative modérée entre le montant emprunté et le taux de versement (-0.27), ce qui reflète une tendance logique : des crédits de faible montant sont souvent remboursés avec des taux plus élevés.

Visualisation Analyse bivariée



Le graphique montre une relation positive entre la durée du crédit et le montant emprunté. La ligne de régression met en évidence une tendance ascendante : les crédits de longue durée sont généralement associés à des montants plus élevés. Cependant, la dispersion des points indique qu'il existe une certaine variabilité autour de cette tendance.

```
Le coefficient de corrélation (numpy) est : 0.62
```

Nous avons également calculé le coefficient de corrélation qui est de 0.62 indique une relation modérée à forte entre la durée du crédit et le montant emprunté. Cela signifie que, bien que la relation ne soit pas parfaite, il existe une tendance significative reliant ces deux variables.

Partie 4 : Analyse multivarié avec l’analyse des composantes principales (ACP)

L'analyse en composantes principales (ACP) est une méthode statistique qui permet de réduire la dimensionnalité des données tout en conservant un maximum d'informations. Elle transforme les variables **quantitatives** initiales en un ensemble de nouvelles variables (appelées **composantes principales**), ordonnées par la quantité de **variance** qu'elles expliquent. Cette méthode est particulièrement utile pour visualiser les relations entre les individus et les variables dans un espace à souvent à deux dimensions. Cette méthode permet de **réduire** les dimensions en conservant le **maximum** d’information. Cela permet d’analyser simultanément des variables. Ainsi, des **relations cachées** peuvent être vue et cela simplifie donc notre analyse.
Ainsi, voici les différentes étapes de l’ACP.

I- La normalisation des données

Dans un premier temps, nous avons scindé les données afin de ne retenir que les variables quantitatives qui sont les seules utilisables pour réaliser l’ACP . Dans un second temps, il est nécessaire de standardiser toutes les variables et de les normaliser en 1 c’est-à-dire pour avoir une moyenne de 0 et une variance de 1. Les variables deviennent des vecteurs normalisés.
L’objectif est de préparer les données dans un format adapté pour réduire les dimensions avec l'ACP, en s'assurant que toutes les variables ont une contribution égale.

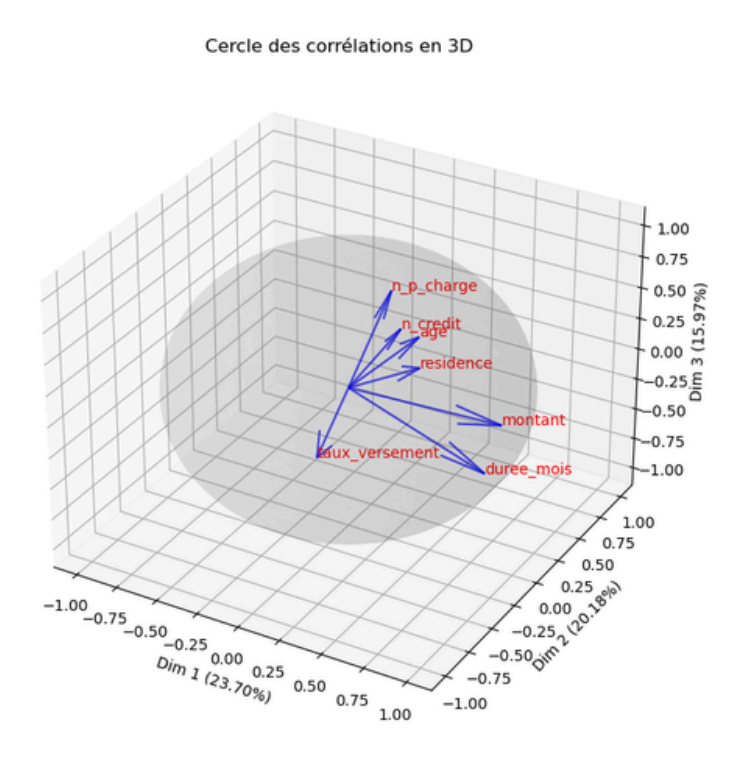
II- Réduction des dimensions

Suite à cela, grâce à une méthode d’algèbre linéaire les données sont décomposés en **vecteurs** propres et en **valeurs** propres. Les **vecteurs propres** décrivent les directions principales dans l'espace multidimensionnel. Les **valeurs propres** quant à eux expliquent la proportion de variance expliquée par chaque composante principale.

Suite à cela, voici le tableau des variances expliquées par chaque dimensions . L'idéale est de sélectionner les dimensions qui à elles seules permettent d'expliquer plus de 50% de la variabilité des variables.

	Dimension	Valeur propre	% Variance expliquée	% cum. var. expliquée
0	Dim1	1.660463	24.0	24.0
1	Dim2	1.413966	20.0	44.0
2	Dim3	1.119302	16.0	60.0
3	Dim4	0.940065	13.0	73.0
4	Dim5	0.869076	12.0	86.0
5	Dim6	0.721383	10.0	96.0
6	Dim7	0.282753	4.0	100.0

Comme on peut le voir sur notre tableau en sorti , il faut donc prendre les 3 premières dimensions afin d'avoir une analyse assez précise et recevable. Cela signifie que l'on se retrouve dans un plan 3D pour l'analyse graphique ce qui rend l'analyse graphique tout de même complexe. (cf code jupyther pour voir plan factoriel 2D).



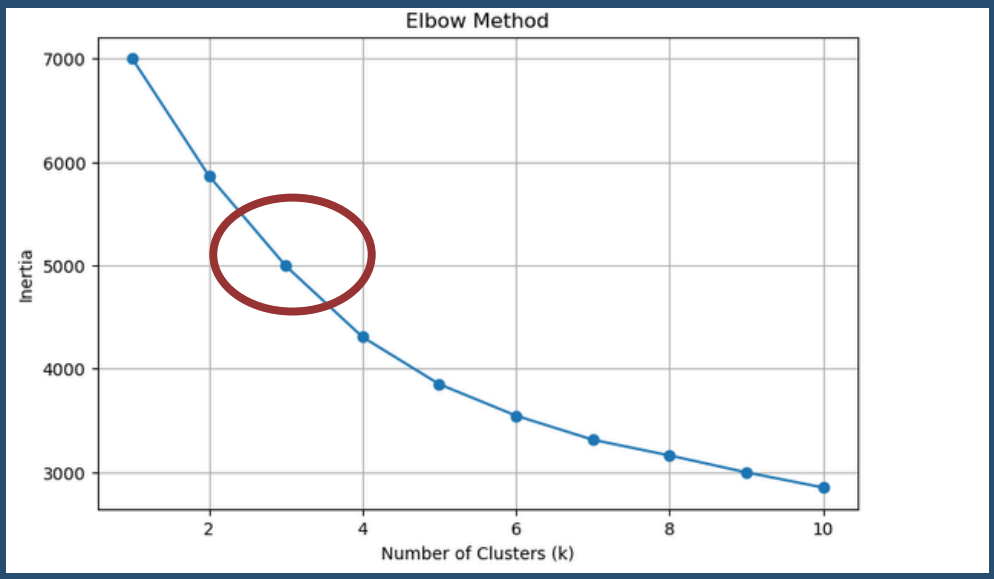
Ce graphique vous permettra de visualiser les variables en fonction des trois premières composantes principales.
La longueur des flèches indique la force de leur corrélation **avec les composantes principales** tandis que leur directions des indiquent comment les variables contribuent à la construction des dimensions.
Les flèches orientées dans une direction proche de l'axe Dim 1 comme montant et duree_mois sont fortement **corrélées** avec cette dimension. Ces variables pourraient expliquer des caractéristiques financières comme la durée des crédits et les montants associés.

Tandis que les variables proches de l'origine (centre du graphique) ne sont pas bien représentées dans les trois premières dimensions, ce qui signifie qu'elles ont une faible corrélation avec ces dimensions.
Le degrés des angles **entre** variables à une signification quant à leur corrélation. L'angle droit signifie qu'il n'y a aucune corrélation alors que plus les angles sont aigus ou obtus plus ils ont une corrélation respectivement positive ou négative. Par exemple, on constate que les variables âges et nombre de crédit sont fortement corrélées positivement.

Méthode k-means

Le clustering K-Means est une méthode utilisée pour regrouper des observations similaires dans des clusters (groupes). Chaque cluster est défini par un centre, calculé comme la moyenne des points qui lui sont attribués. Cette méthode est utile pour explorer des ensembles de données complexes et identifier des structures sous-jacentes.
Dans cette analyse, nous avons utilisé la méthode Elbow pour déterminer le nombre optimal de clusters (kkk), puis appliqué K-Means avec k=3k=3k=3 pour segmenter les données. Les résultats ont été visualisés à l'aide d'une réduction dimensionnelle via l'ACP.

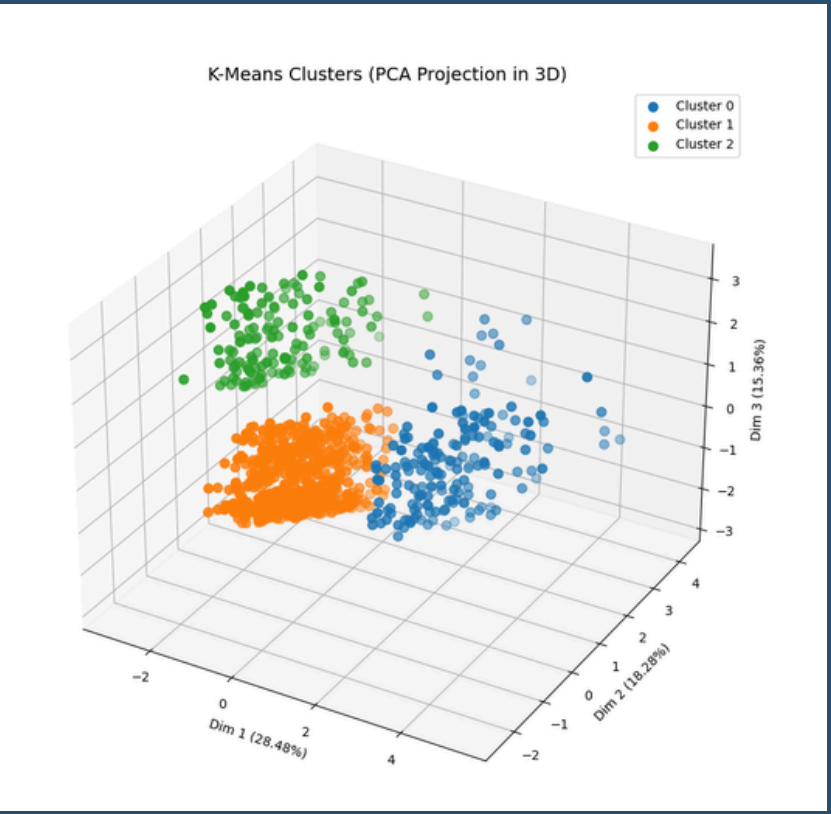
Méthode Elbow :



Ce premier graphique montre la méthode Elbow, utilisée pour **déterminer le nombre optimal de clusters (kkk)** en fonction de l'inertie. L'inertie représente la somme des **distances** au carré entre les points d'un cluster et leur **centre**.
L'inertie diminue lorsque le nombre de clusters augmente, car les groupes deviennent plus spécifiques.
On observe un point de coude pour k=3. C'est à ce niveau que l'inertie commence à diminuer **moins rapidement**. Ce point indique le nombre optimal de clusters, car il représente un bon compromis entre complexité et efficacité de la segmentation. Dans notre graphiques de K-means, on choisira donc 3 groupes.

Méthode k-means

Visualisation des clusters avec PCA :



A partir de la des résultat de la réduction dimensionelle faite plus haut avec l'ACP, La visualisation des données des clustering est projetté un espace tridimensionnel tout en conservant une grande partie de la variance.
Cesecond graphique illustre les clusters identifiés par K-Means sur ces trois premières dimensions :

- Dim 1 explique 23,70% de la variance des données.
- Dim 2 explique 20,18% de la variance.
- Dim 3 explique 15,36% de la variance.

Ainsi, ces trois dimensions permettent d'expliquer plus de 60% de la variance des données.

Méthode k-means

Chaque point représente une observation, et sa couleur correspond au cluster auquel elle appartient. Les clusters sont bien séparés, ce qui montre que **l'algorithme** a efficacement regroupé les observations similaires. Cependant, un léger chevauchement entre certains groupes peut être observé, probablement dû à des similarités intrinsèques entre certaines observations ou à une perte d'information causée par la réduction dimensionnelle.

L'ajout des clusters aux données originales permet d'examiner leurs caractéristiques distinctives (non présentées ici). En général, ces clusters peuvent révéler des groupes naturels ou des catégories spécifiques au sein des données, utiles pour identifier des tendances ou des comportements spécifiques et pour orienter des actions (segmentation de clientèle, ciblage marketing...).

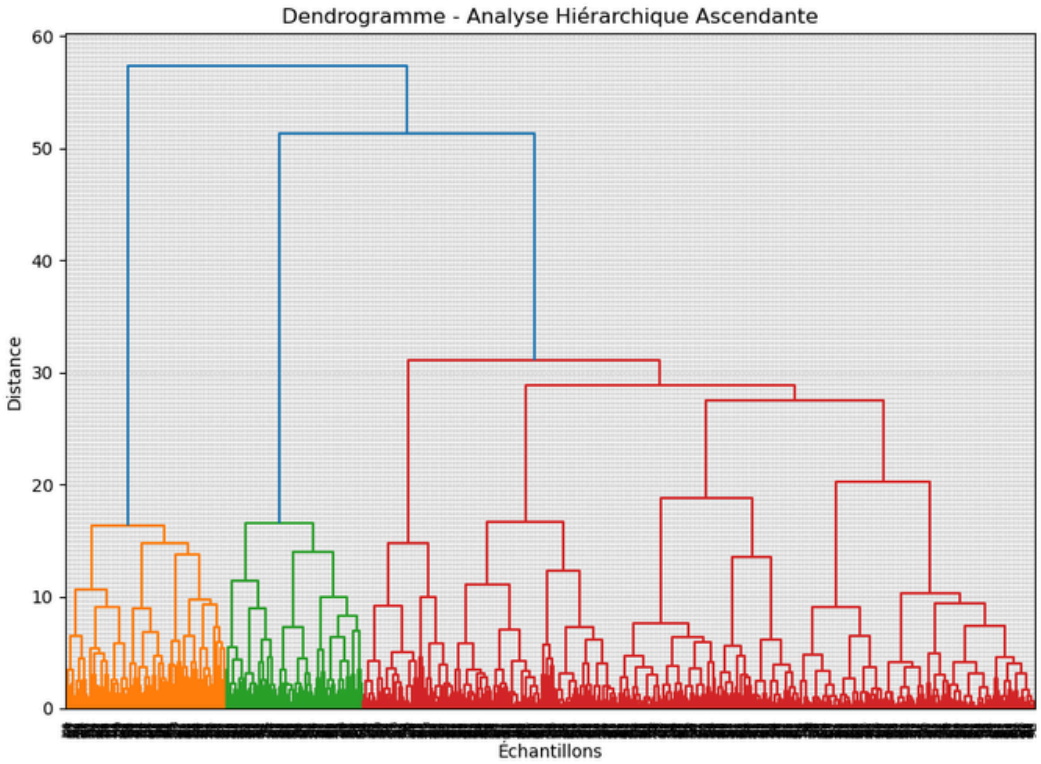
Cette analyse a montré que la méthode Elbow suggère k=3k=3k=3 comme le nombre optimal de clusters L'algorithme K-Means a segmenté efficacement les données, avec une bonne séparation visible dans la projection ACP.

Ces résultats peuvent servir de point de départ pour une exploration plus approfondie des clusters. Par exemple, il serait pertinent d'analyser les caractéristiques propres à chaque cluster pour mieux comprendre les différences entre eux et les exploiter dans un contexte spécifique (par exemple, en recherche, en marketing ou en analyse décisionnelle).

Ici, **Cluster 1** représente majoritairement des jeunes adultes avec des crédits de faible montant et courte durée.
Le **Cluster 2** regroupant des clients intermédiaires, avec des crédits modérés et un risque équilibré.
Et enfin le **Cluster 3** avec les Clients "à risque" avec des crédits élevés, longue durée, et dispersion importante des montants.
Le cluster 3 est donc celui d'intérêt puisque grâce aux spécificité de ceux qui le compose nous pouvons faire un portrait des client ciblé à risque et donc prendre des précaussions.

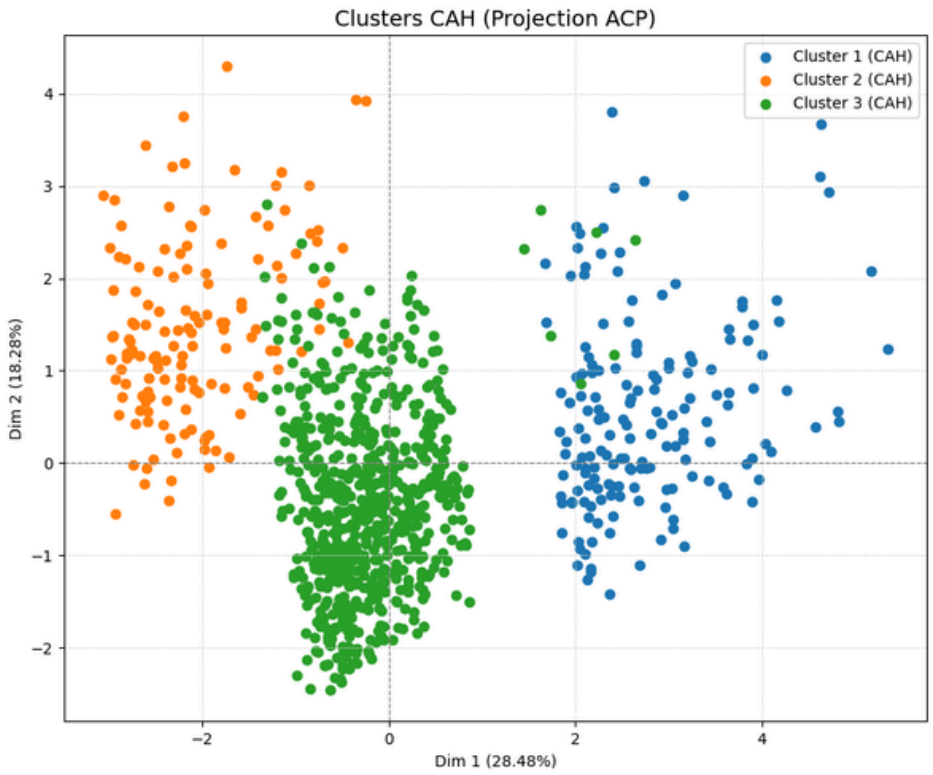
Méthode CAH

En complément des méthodes d'analyse exploratoire telles que l'ACP et le K-Means, l'Analyse Hiérarchique Ascendante (CAH) a été utilisée pour identifier des regroupements au sein des données. Contrairement au K-Means, qui nécessite de définir un nombre de clusters au préalable, la CAH permet de visualiser les regroupements sous forme d'un dendrogramme et de choisir un nombre optimal de clusters en fonction de la distance de coupure.



Le dendrogramme révèle une structure hiérarchique des données, avec une séparation claire entre les groupes. En fixant la distance de coupure (indiquée par une ligne horizontale), nous avons identifié 3 clusters principaux, cohérents avec les résultats obtenus par le K-Means.

Méthode CAH



Les clusters identifiés par la CAH ont été projetés sur les dimensions principales issues de l'ACP pour faciliter leur visualisation. Cette projection montre une bonne séparation entre les groupes, ce qui confirme la qualité des regroupements. Chaque cluster possède des caractéristiques distinctes dans l'espace factoriel, renforçant l'idée de sous-populations homogènes dans les données.

En complément de l'ACP et du K-Means, l'Analyse Hiérarchique Ascendante (CAH) a permis d'identifier des groupes similaires au sein des données tout en offrant une visualisation structurée via un dendrogramme. Cette méthode constitue une alternative pertinente pour explorer la hiérarchie et les relations entre les individus, apportant une dimension supplémentaire à l'interprétation des clusters obtenus.

Conclusion

L’analyse menée sur les données bancaires relatives aux crédits allemands a permis de mettre en lumière des éléments clés pour différencier les clients à faible et à haut risque.

Avec l’**analyse descriptive et bivariée** certaines relation et variables ont retenue notre attention et nous avons donc poussé plus loin la réflexion avec une étude multivarié suivant 3 méthode différentes.
De plus la corrélation entre la durée du crédit et le montant emprunté est modérée à forte (0.62), confirmant que les crédits de longue durée sont souvent plus élevés.
En revanche, la variable cible (bon ou mauvais risque) présente une corrélation relativement faible avec les autres variables (< 0.3), indiquant que plusieurs facteurs combinés influencent le risque de crédit.

Avec l’ **ACP** (**Analyse en Composantes Principales**) ,les trois premières dimensions expliquent 60 % de la variance totale des données, rendant possible une visualisation pertinente des variables et individus. Les variables telles que montant emprunté et durée du crédit sont fortement corrélées à la première dimension, suggérant leur poids dans l'analyse des comportements financiers.
L’algorithme du **Clustering K-Means** quant à lui à subdiviser la population en 3 et pour cibler les éléments à risque une analyse approfondit doit être faite sur le cluster 3.
Enfin, avec le **CAH** (Classification Ascendante Hiérarchique), le dendrogramme confirme également **3 regroupements principaux**, cohérents avec ceux identifiés par K-Means. Cette approche hiérarchique met en évidence des relations plus détaillées entre les individus, notamment une segmentation plus fine des "clients intermédiaires".

Les analyses combinées offrent une compréhension approfondie des comportements financiers des emprunteurs. Les clients "à risque" se distinguent par des crédits élevés et de longue durée, tandis que les bons clients présentent des profils plus équilibrés, avec des crédits généralement plus courts et des montants modérés.
En croisant les résultats des méthodes ACP, K-Means, et CAH, nous avons pu non seulement valider la segmentation en trois groupes, mais également proposer des axes d'amélioration pour la gestion des risques bancaires. Par exemple, un focus particulier sur les crédits de longue durée et un suivi renforcé des emprunteurs jeunes avec des montants élevés pourraient permettre aux banques d'anticiper et de réduire les défauts de paiement.
Cette analyse constitue une base solide pour le développement de modèles prédictifs et une prise de décision éclairée dans la gestion des crédits bancaires.