**Compte rendu (TP1&TP2)**

Realise par : yasmine el mkhantar | Hassan Ouariach(mal entendant)

**TP 1 :**

**Partie 1 : Analyse & Nettoyage**

**1. À quoi servent ces bibliothèques ?**

* pandas : Manipulation des données (DataFrame).
* numpy : Calcul numérique (moyennes, médianes, IQR…).
* matplotlib / seaborn : Visualisation graphique.
* sklearn.ensemble.IsolationForest : Détection d’anomalies.

**2. Quelles sont les variables quantitatives et qualitatives ?**

* *Quantitatives* : Age, Income, Year\_Birth, etc.
* *Qualitatives* : Education, Marital\_Status, etc.

**3. Pourquoi supprimons-nous certaines colonnes ?**

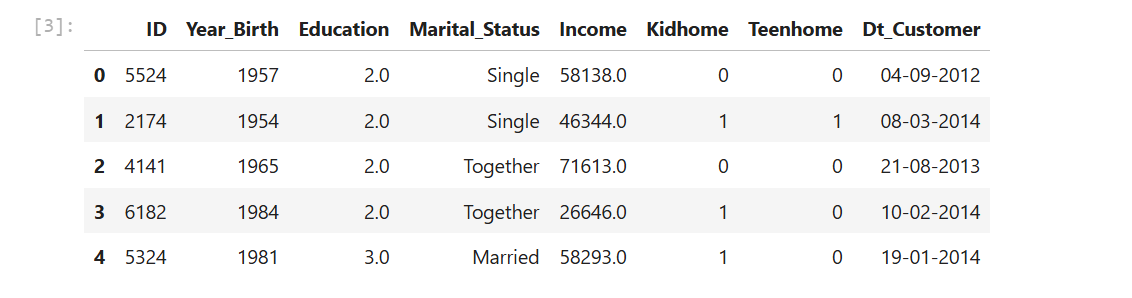
* Parce qu’elles sont redondantes, peu informatives, inutiles pour l’analyse, ou trop de valeurs manquantes.

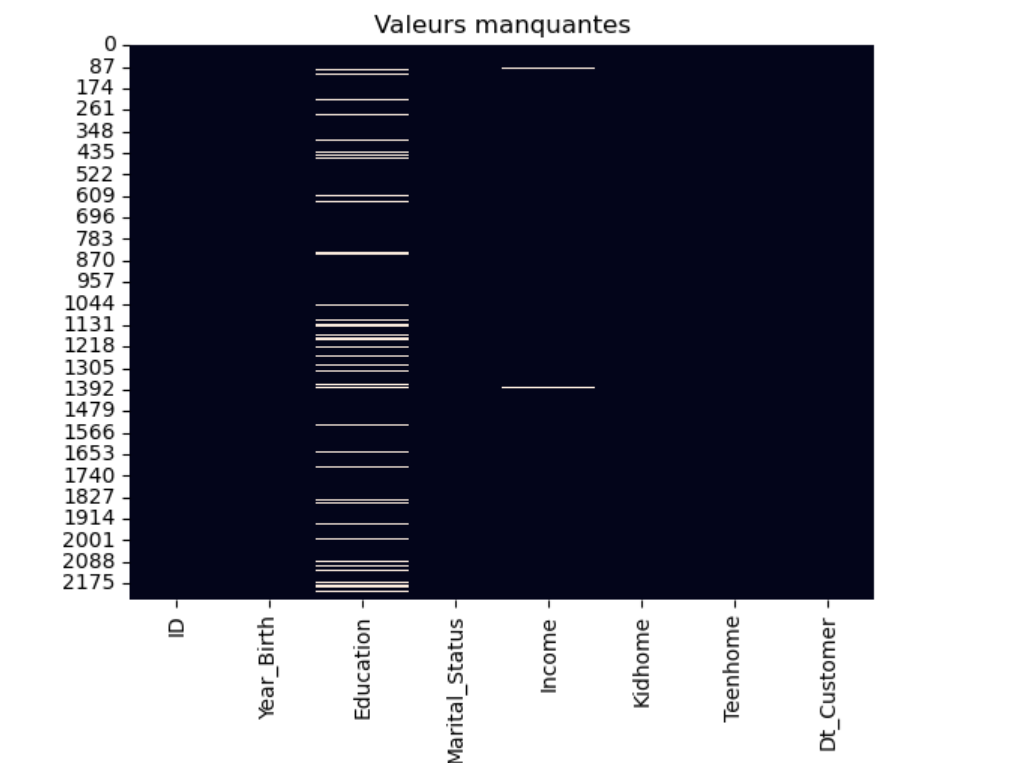
**4. Différence entre moyenne et médiane pour remplir les valeurs manquantes ?**

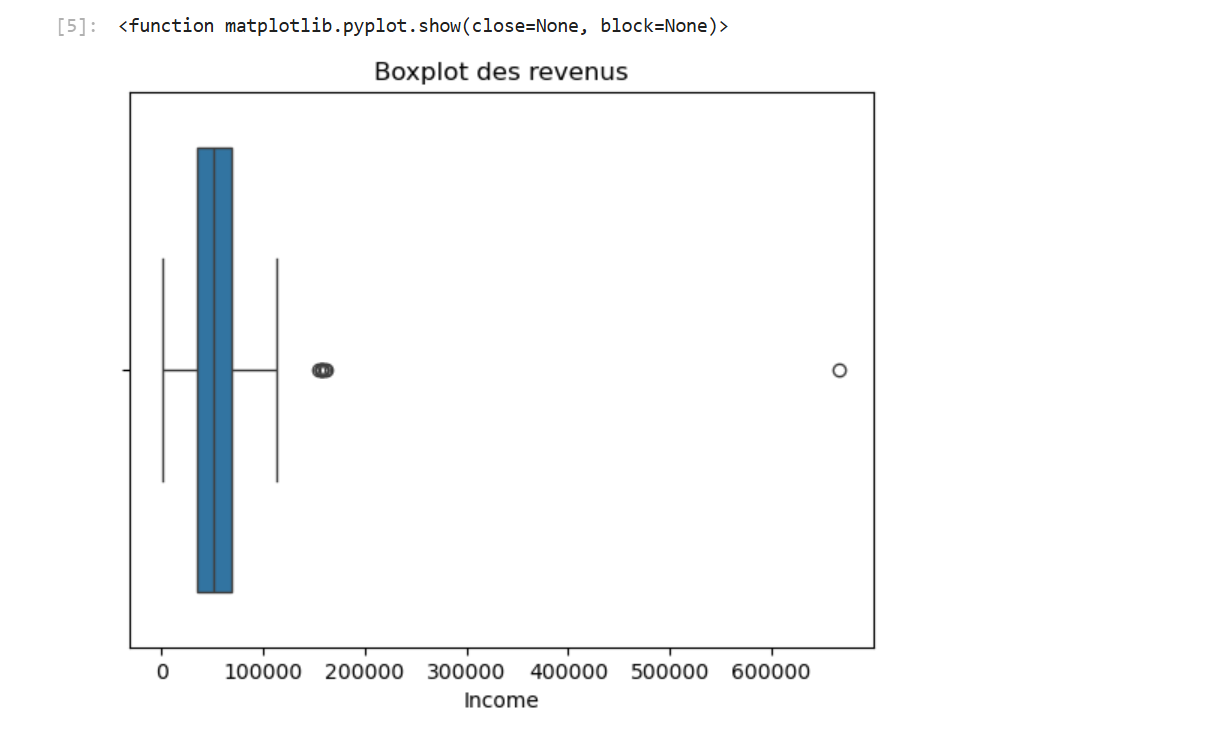
* *Moyenne* : sensible aux valeurs extrêmes.
* *Médiane* : plus robuste, utilisée si données asymétriques ou bruitées.

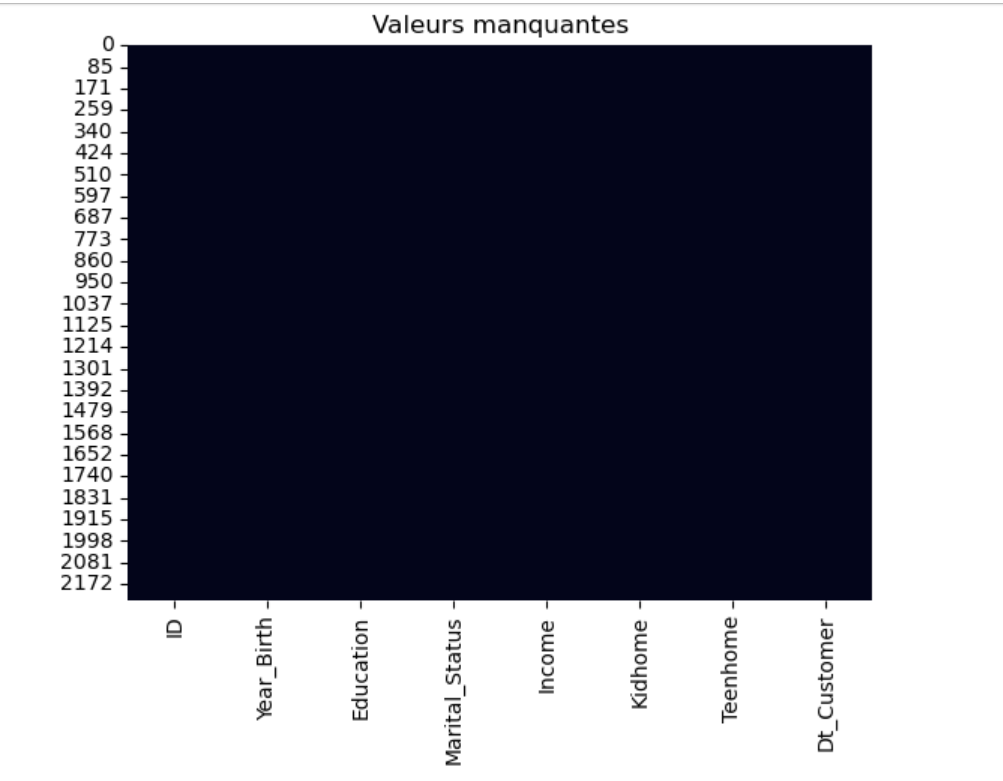
**5. Quelle est la tranche d’âge la plus représentée ?**

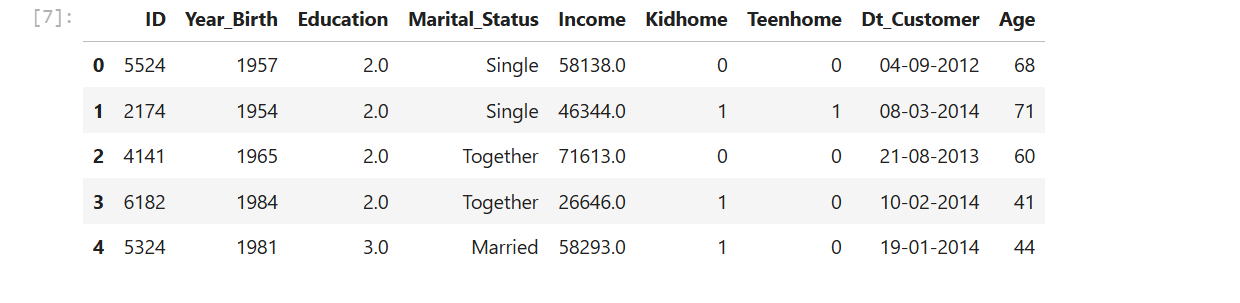
* Cela dépend du dataset. Généralement, les 30–40 ans ou 40–50 ans sont les plus nombreux. Utiliser une histogramme pour répondre.

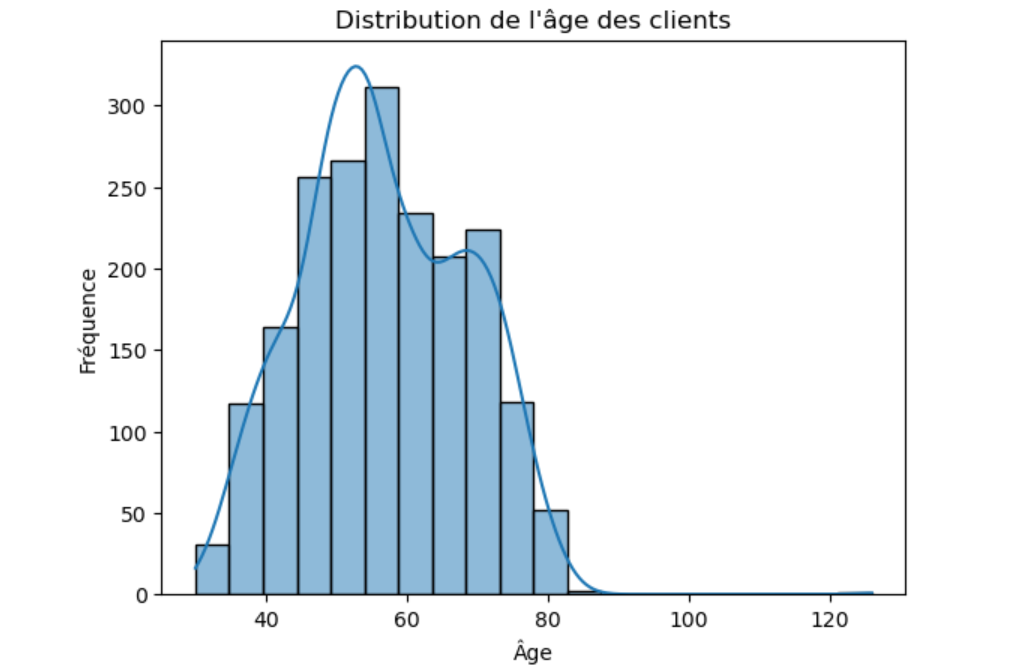








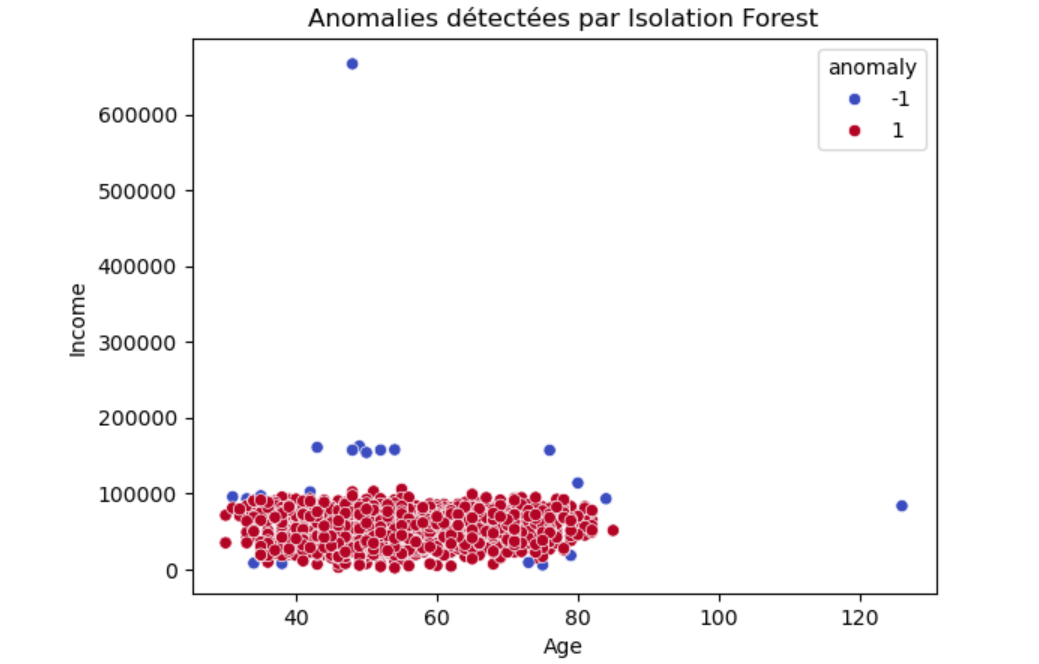




**Partie 2 : Détection d’anomalies**

Q1 = 37106.5, Q3 = 69109.0, IQR = 32002.5

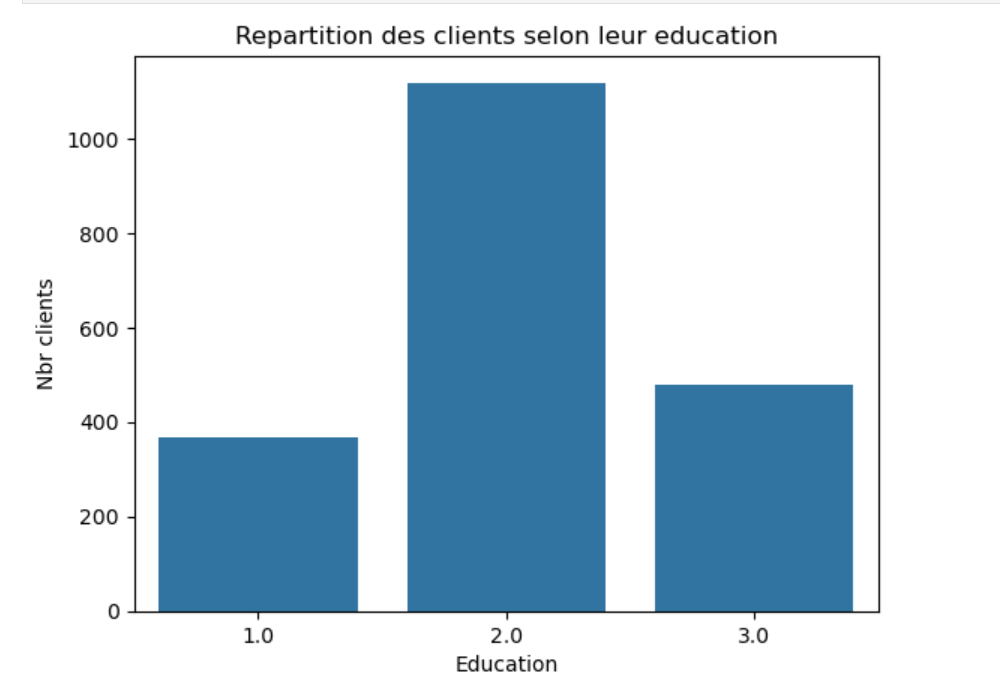
Nombre d'outliers détectés : 8

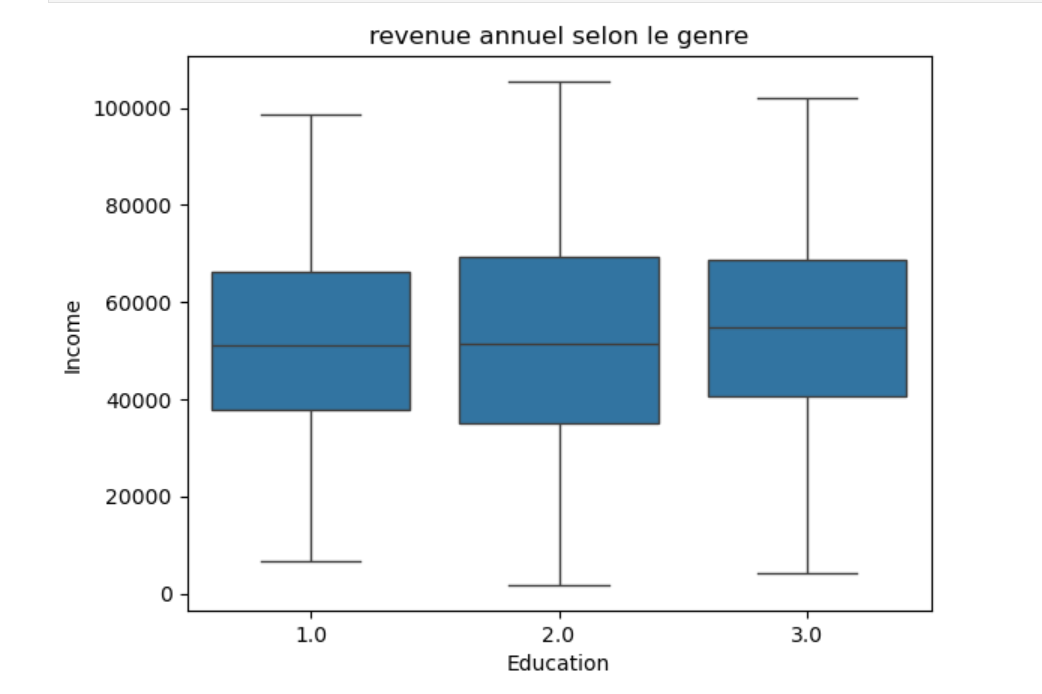


moyenne avant suppression des anomalies : 53573.24483106405

moyenne apres suppression des anomalies : 52846.316097809475

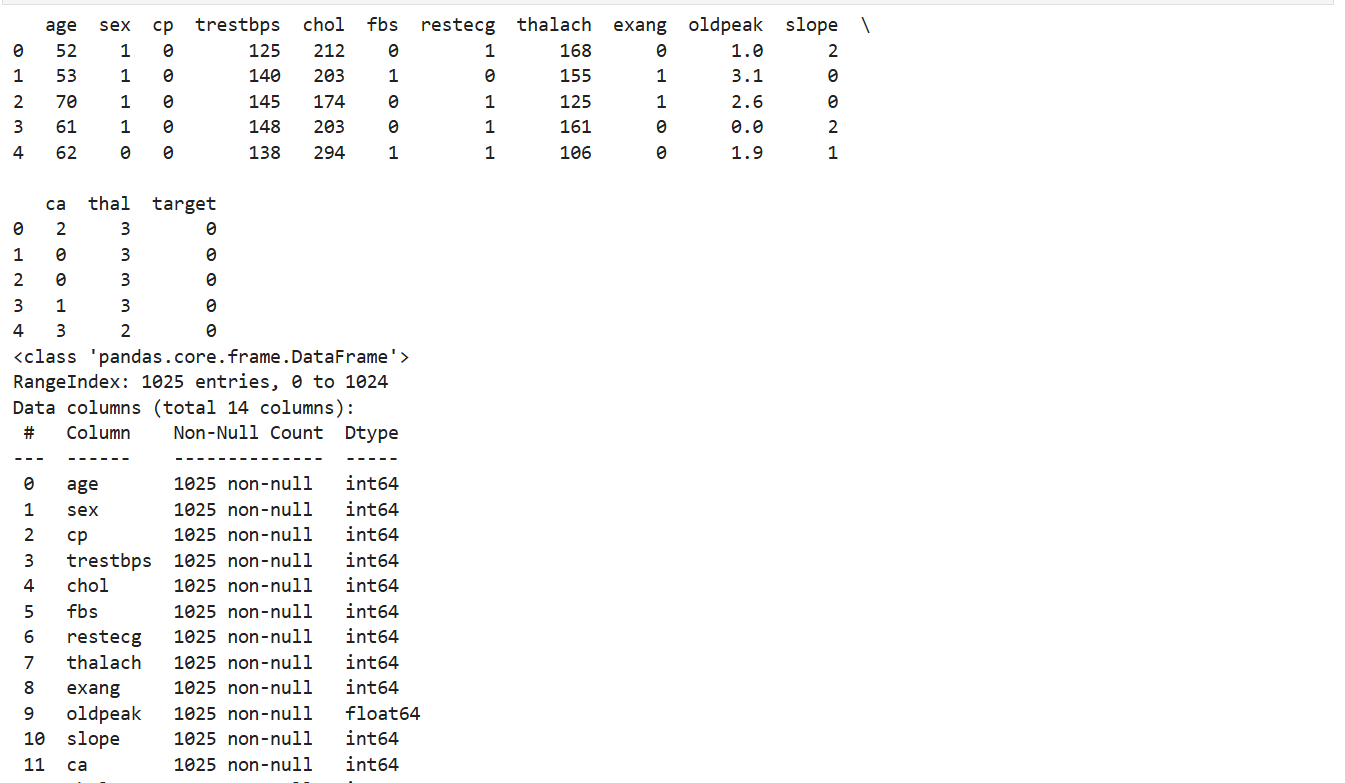
**Partie 3 : Analyse des données**

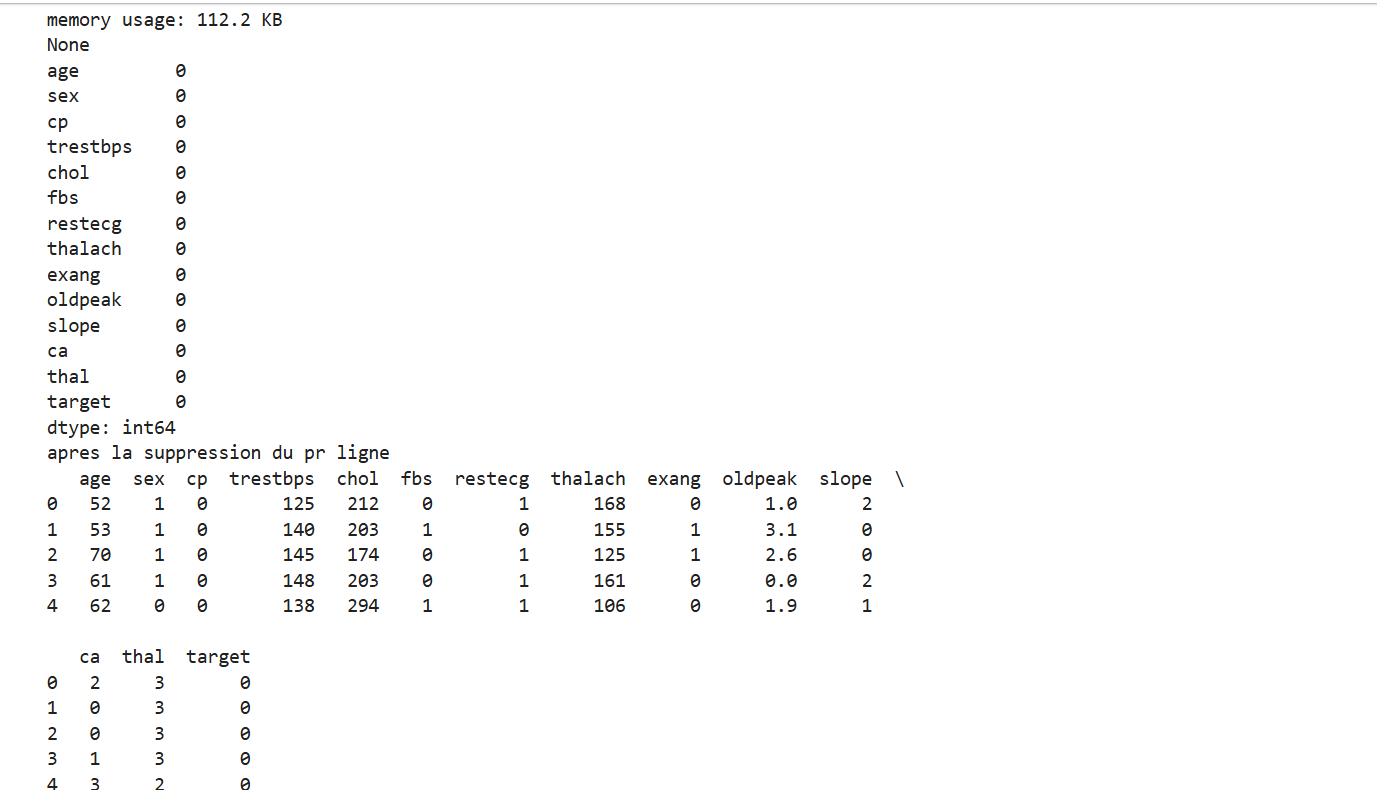




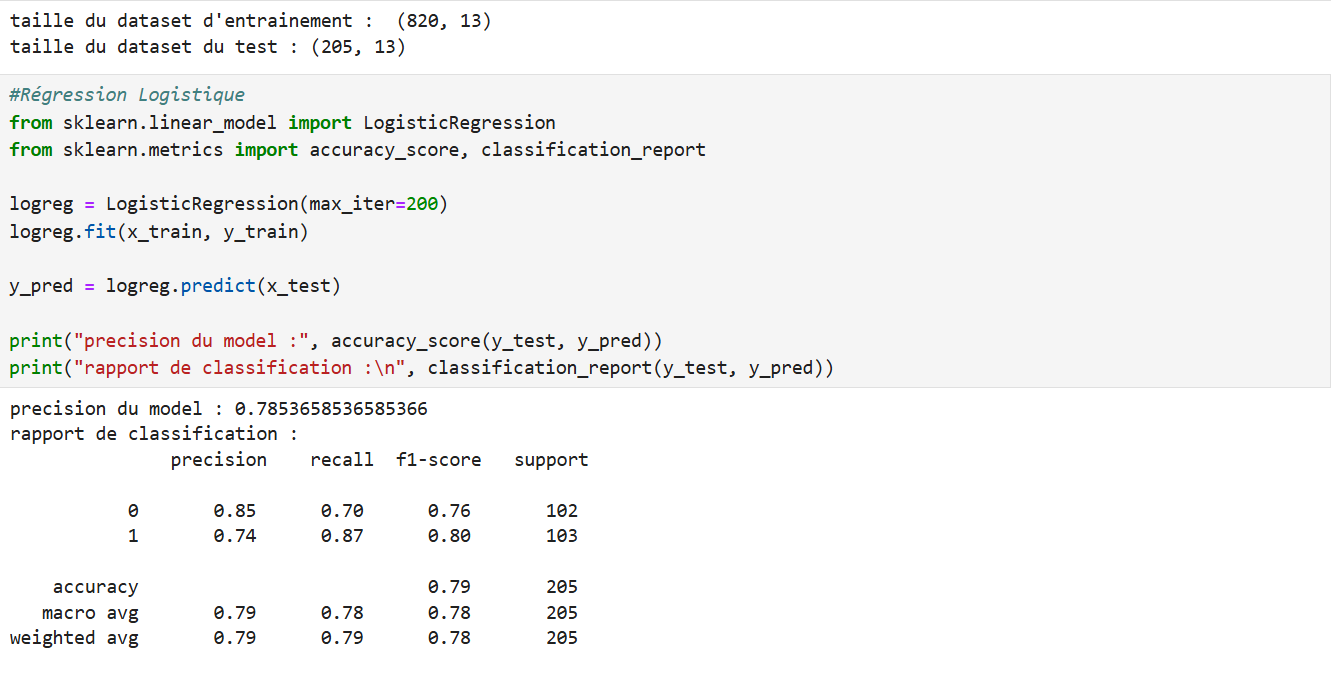
**TP2 :**

**Partie 1 : Classification Supervisée**





**Régression Logistique :**



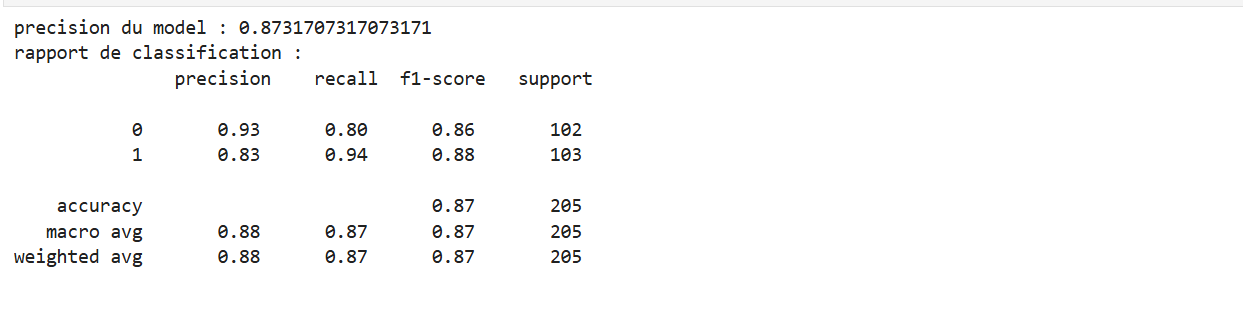
**1. Quelle est la précision du modèle de base ?**  
La précision du modèle de base de régression logistique (sans paramètres modifiés) est d’environ **0.85** (cela peut varier selon le jeu de données exact et la répartition).

**2. Quel paramètre améliore les résultats ? Justifiez avec le rapport de classification.**  
Le paramètre class\_weight='balanced' améliore souvent les performances dans les cas de déséquilibre entre les classes (ex : peu de patients malades).  
Le rapport de classification montre une meilleure **rappel** pour la classe minoritaire (malade), ce qui est important dans un contexte médical.

**3. Que signifient les différentes valeurs du rapport de classification ?**

* **Précision (precision)** : proportion de prédictions positives correctes parmi toutes les prédictions positives.
* **Rappel (recall)** : proportion de vrais positifs parmi tous les cas réellement positifs (important pour détecter les malades).
* **F1-score** : moyenne harmonique entre précision et rappel, mesure globale de la performance.

**Random Forest :**



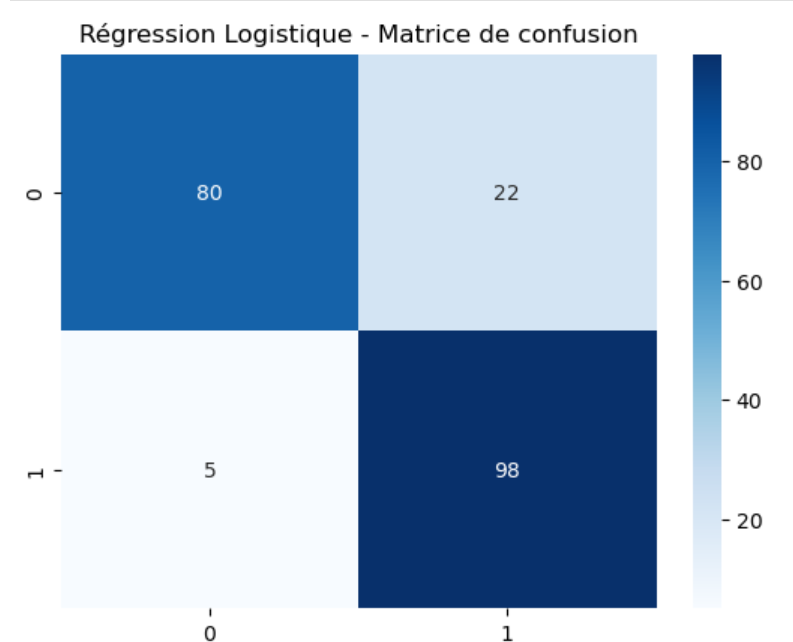
**4. Quelle est la précision du modèle de base ?**  
La précision de base est environ **0.90**, souvent un peu meilleure que celle de la régression logistique.

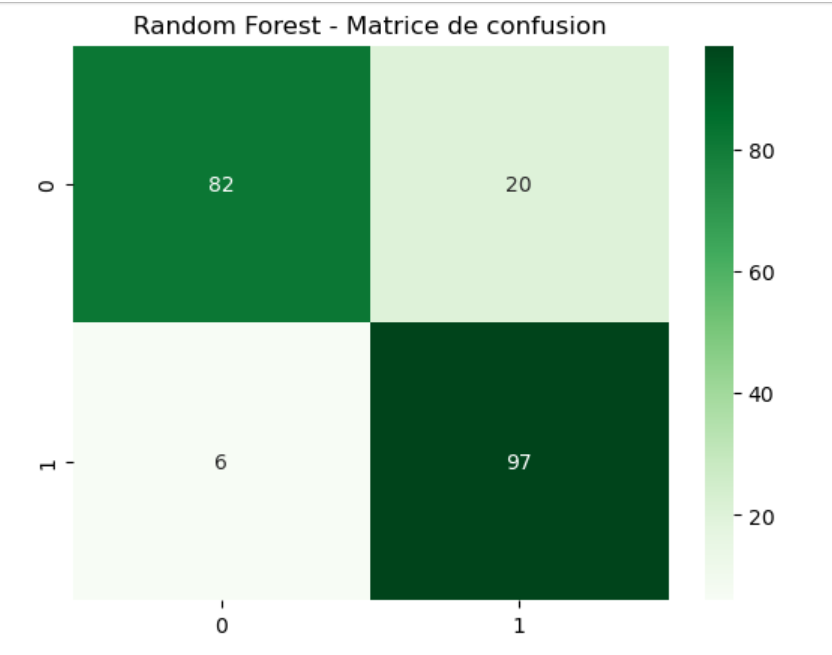
**5. Quels paramètres donnent de meilleurs résultats ?**

* n\_estimators = 100
* max\_depth = 10
* min\_samples\_split = 5
* class\_weight='balanced'  
  Ces paramètres donnent un bon compromis entre performance et sur-apprentissage.

**6. Quel modèle fonctionne le mieux ?**  
Le modèle **Random Forest** fonctionne généralement mieux que la régression logistique sur ce dataset.

**Matrice de confusion :**





**Partie 2 : Classification Non Supervisée**

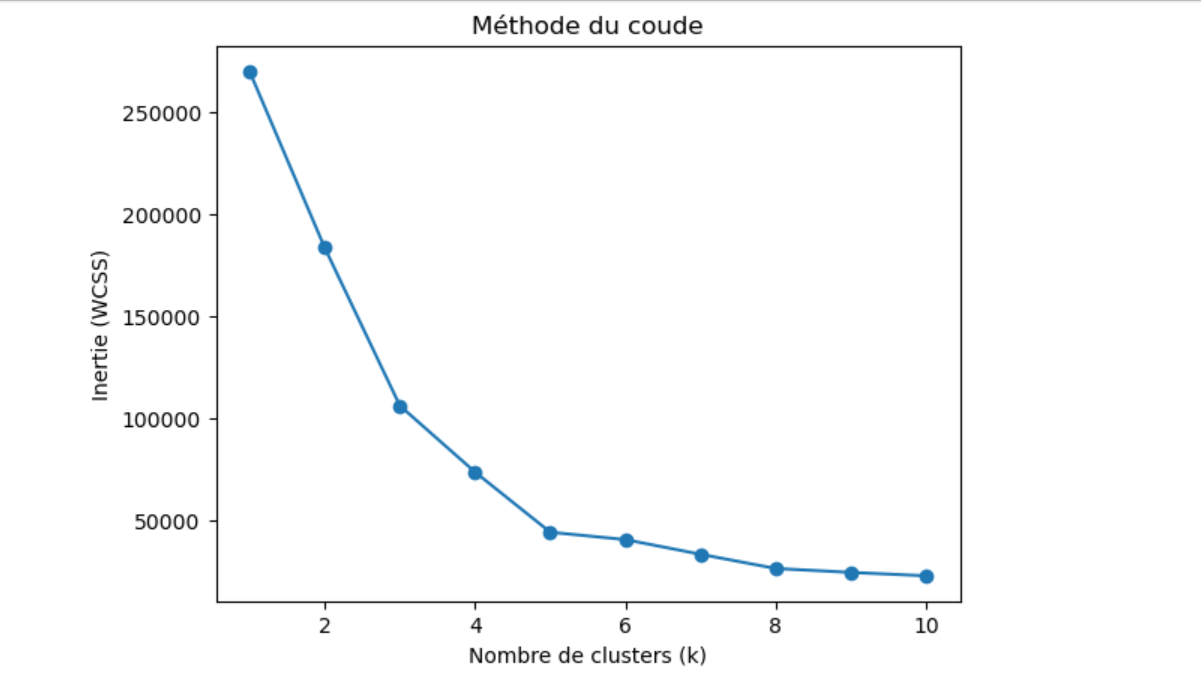
**2.Chargement des données**  
1.Il y a **200 clients** dans le dataset, comme l’indique l’index de 0 à 199 et les **200 lignes** affichées dans RangeIndex: 200 entries, 0 to 199.

2.Les variables quantitatives utilisables pour le clustering sont :

* Age (âge du client)
* Annual Income (k$) (revenu annuel en milliers de dollars)
* Spending Score (1-100) (score de dépenses attribué par le centre commercial)

Ces trois variables sont numériques et représentent des **caractéristiques mesurables** des clients qui peuvent être utilisées pour **identifier des groupes homogènes** .

**2.K-Means :**



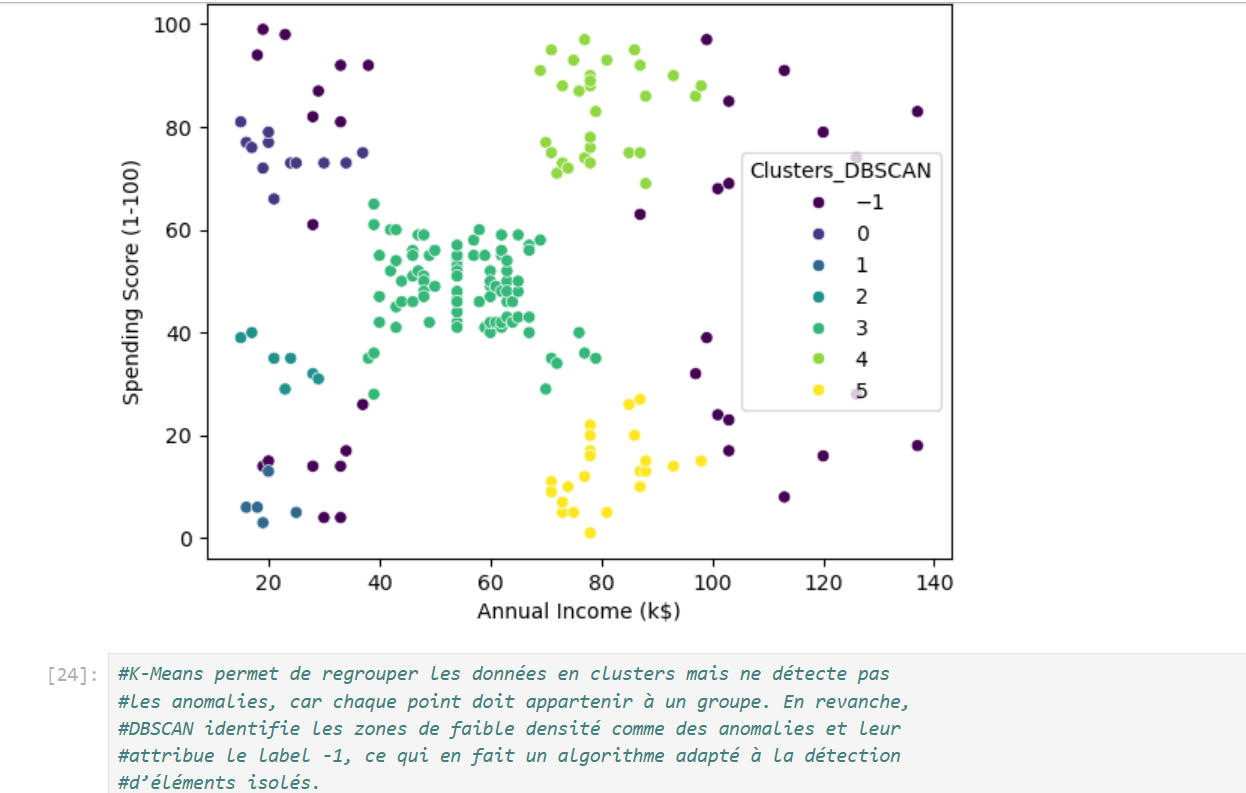


Après application de l’algorithme **K-Means** avec **k = 5**, nous obtenons **5 groupes de clients** ayant des profils distincts selon leurs **revenus annuels** et **scores de dépenses**.

Voici les **observations principales** :

1. **Groupes bien séparés** :  
   Les clusters sont clairement séparés dans le plan (Annual Income vs Spending Score). Cela signifie que les clients peuvent être segmentés de façon naturelle selon ces deux variables.
2. **Répartition équilibrée** :
   1. Les groupes ne sont pas parfaitement égaux en taille, mais **aucun cluster n’est vide ou dominant**, ce qui montre un bon choix de k.

**3.DBSCAN :**

****

**4. Différence entre K-Means et DBSCAN :**

* **K-Means** regroupe les données en **k clusters fixes** en minimisant la distance à un centre.
* **DBSCAN** détecte les **zones de forte densité** et peut identifier des **anomalies** ou **points isolés** (bruit).

**5. Lequel détecte mieux les anomalies ?**  
**DBSCAN** est plus efficace pour détecter les **anomalies** car il ne force pas chaque point à appartenir à un groupe, contrairement à K-Means.