TP 03

Objectifs :
— librairie pandas : manipulation et agrégation
— visualisation des données
— nettoyage de données

Exercice 1 : Demandes d'emprunts et crédits

Nous allons travailler sur un fichier de prêts immobiliers. Chaque ligne de ce fichier correspond à un prêt qui a été accordé à un client. Chaque client est identifié de manière unique par son identifiant et nous disposons des informations associées suivantes :
— la ville et le code postale de l'agence où le client a contracté le prêt
— le revenu mensuel du client
— les mensualités remboursées par le client
— la durée du prêt contracté, en nombre de mois
— le type de prêt
— le taux d'intérêt

Votre rôle cette fois-ci va être de modifier ce jeu de données pour calculer différentes variables nécessaires pour identifier les clients qui sont à la limite de leur capacité de remboursement et déterminer les bénéfices réalisées par la banque.

Q1. Importez les données du fichier "`data_mm03_loans.csv`".

Q2. Commençons par enrichir les données.

Q2.1. Créez une nouvelle variable `taux_endettement` correspondant au pourcentage du revenu remboursé par mois pour chaque individu. Vous arrondirez le résultat à 2 chiffres après la virgule

Q2.2. Pour éviter toute confusion, renommez la variable `taux` en `taux_interet` (via la méthode `rename(...)`).

Q2.3. Créez une variable `cout_total` qui correspond au coût total comme son nom l'indique, à partir des mensualités de remboursement et de la durée.

Q2.4. Créez enfin une variable `benefices` correspondant aux bénéfices mensuels réalisés par la banque sur le prêt. On simplifiera le calcul des bénéfices par :

$$benefices = \frac{cout\_total \times taux\_interet}{100 \times 24}$$

Q3. Un peu de visualisation préliminaire.

Q3.1. En utilisant la fonction `scatter(...)` de `matplotlib.pyplot`, tracez les nuages de points des revenus des clients par rapport aux taux d'intérêt de leur prêts, pour les agences de Paris et Toulouse uniquement. Il faut pouvoir distinguer les prêts par agence. N'oubliez pas la légende et le titre.

Q3.2. Réalisez le même graphique pour toutes les agences. Votre code doit fonctionner quel que soit le nombre d'agence (et donc l'usage d'une boucle devient inévitable).

Q3.3. Au lieu d'utiliser `matplotlib`, refaites plus simplement ce même graphique grâce à `seaborn`, sa fonction `scatterplot` et l'option `hue`.

Q4. Un peu de filtrage et de manipulations.

Q4.1. Affichez les prêts de type '`automobile`'.

Q4.2. Affichez uniquement les identifiants des clients ayant contractés un prêt de type '`automobile`'.

Q4.3. Affichez les identifiants et les revenus des clients ayant contractés un prêt de type '`automobile`'.

Q4.4. Nous faisons un peu de prospection. On ne propose de prêts qu'aux clients capable de rembourser ! Affichez les clients ayant un revenu supérieur à 4000 € mensuel.

Q4.5. Affichez les personnes ayant un revenu supérieur à 4000 € mensuel et ayant déjà contracté un prêt automobile.

Q4.6. Via un tri (`.sort_values(...)`), affichez les 5 prêts les plus rentables pour la banque. Quels sont les types de prêts les plus rentables ?

Q4.7. Fixez le taux d'intérêt à 1.05% pour les trois premiers clients du *dataframe*. Fixez celui des prêts du client d'identifiant 7 à 1.25%.

Q4.8. Dess erreurs se sont glissées dans les taux des prêts réalisés par l'agence parisienne et par l'agence marseillaise. Affectez un taux unique fixe de 1.33 pour les crédits de l'agence de Paris et augmentez le taux d'intérêt des prêts de l'agence de Marseille de 0.5%.

Q5. Quelques agrégations.

Q5.1. Combien de prêts automobiles ont été accordés ? Quel est le coût total moyen de ces derniers ?

Q5.2. En utilisant `matplotlib`, représentez un diagramme à barres (`plt.bar`) illustrant le nombre de prêts selon leur type. Représentez également un diagramme circulaire (`plt.pie`) représentant les proportions de prêts selon leur type.

Q5.3. Générez un diagramme à barres représentant le nombre de prêts accordés par type et par agence.

Q5.4. Le taux d'endettement maximal légalement autorisé est de 35%. Combien de clients ont dépassés ce seuil ? Qu'en est-il si on considère uniquement pour l'agence de Paris ?

Q5.5. Cette information nous semble particulièrement intéressante et nous souhaitons la sauvegarder. Créez une nouvelle colonne booléenne "`risque`" indiquant si le client est un client dit "*à risque*".

Q5.6. On peut remarquer que certains clients ont contracté plusieurs prêts au sein de notre établissement. Cela fausse donc potentiellement les calculs réalisés précédemment. Créez un *dataframe* `profil_clients` où il y n'y a qu'une seule ligne par client, avec le résumé de ses informations (somme des remboursements, du taux d'endettement, du coût total et des bénéfices réalisés.)

Q5.7. Recalculez le nombre exact de personnes en situation bancaire risquée à partir du taux d'endettement (qui doit être supérieur à 35%).

Q5.8. Quel est le bénéfice mensuel total réalisé par l'agence toulousaine ?

Q5.9. Calculez le bénéfice dégagé par chacune des agences, par types de prêts. Vous présenterez vos résultats sous la forme d'un tableau simple (via un `group_by`)

Q5.10. Pour aller plus loin, on souhaite avoir un tableau à double entrée (via la méthode `pivot_table(...)`) présentant cette fois-ci les bénéfices moyens réalisés par chaque agence, pour chaque type de prêt. Quelle ville semble la plus intéressante où développer les prêts immobiliers ?

Q5.11. Représentez le bénéfice mensuel réalisé en fonction du revenu du client pour les prêts immobiliers en utilisant des couleurs différentes pour les agences (via un `scatterplot(...)`).

Q5.12. Générez un diagramme à barres (`barplot(...)`) illustrant les remboursements mensuels par agence et par type de prêts (via les couleurs).

Exercice 2 : Les jeux olympiques

Q1. Importation et exploration des valeurs nulles.

Q1.1. Importez le fichier CSV sur les jeux olympiques dans un *dataframe*, en utilisant toutes les colonnes et sans spécifier d'index.

Q1.2. Combien y-a-t-il de valeurs nulles par colonnes ? et au total ?

Q2. Cherchons à identifier les données non-significatives.

Q2.1. On cherche à identifier si certains sports ne sont pas significatifs. Combien de valeurs uniques y-a-t-il sur les sports ? Comptez le nombre d'occurrences de chacun et affichez les 20 sports les moins représentés, avec le nombre d'occurences.

Q2.2. Y-a-t-il des pays qui n'apparaissent moins de 10 fois dans l'histoire des jeux olympiques (selon ces données). Lesquels ? Quels sont les trois pays les plus représentés ?

Q3. Posons-nous des questions un peu moins immédiates.

Q3.1. Quel était le ou la plus vieil(le) athèlte à avoir participé aux jeux olympiques ? À quel âge ? Dans quel discipline ? A-t-il eu une médaille ? Trouver le ou la plus vieil(le) athelète médaillé dans l'histoire des jeux le cas échéant.

Q3.2. Dans quel jeu y-a-t-il eu le plus d'athlètes inscrits ?

Q4. Tracez l'évolution du nombre de médailles au cours des jeux, en séparant jeux d'hiver et jeux d'été.

Q5. Cherchons à caractériser certaines distributions.

Q5.1. Générez une figure avec trois boîtes à moustaches, si possible côte à côte dans une même figure, pour les âges, les tailles et les poids. Faites apparaître les outliers.

Q5.2. Caractérisez les distributions d'âge, de taille et de poids avec les valeurs moyenne et médiane, la variance, l'asymétrie et l'aplatissement.

Q5.3. Parmi ces 3 variables, laquelle s'étale le plus (à droite ou à gauche) ? Laquelle est la plus centrée ? Laquelle a le plus d'amplitude ?

Q5.4. Tracez les distributions empiriques des âges, des tailles et des poids, si possible côte-à-côte sur la même figure, et vérifiez vos assertions.

Q5.5. Parmi ces 3 variables, entre lesquelles y-a-t-il la plus grande corrélation ?