TD 2

Prise en main de Numpy, Pandas et Matplotlib

Lors de ce TD, vous vous familiariserez avec trois packages incontournables de python, en particulier pour les praticiens de la science des données et des méthodes d'intelligence artificielle. Vous avez à votre disposition :

- Une *cheat-sheet* (aide-mémoire) de ces frameworks. En cas d'examen, il s'agira de la seule ressource autorisée, en plus de vos notes manuscrites.
- Un environnement de développement python avec NumPy, Pandas et Matplotlib fonctionnels. Vous utiliserez cet environnement pour pratiquer.

Servez-vous des questions ci-dessous pour parfaire votre compréhension de la manipulation de données sous python.

Thématique 1 (NumPy)

1. Créez un tableau NumPy 1-D de taille 20, nommé **heights**, initialisé avec la valeur -1.
2. Affichez le type des valeurs contenues dans le tableau.
3. Créez une fonction afin de saisir 20 tailles humaines dans le tableau. Les tailles seront saisies en mètres.
4. Calculez la moyenne des éléments que vous avez saisis dans une variable **mean_height.**
5. Dans un nouveau tableau **spread**, calculez l'écart à la moyenne de chaque élément.
6. Affichez l'indice de l'élément qui a le plus grand écart à la moyenne.
7. Sauvegardez les tableaux **heights** et **spread** sur le disque.
8. Chargez les deux tableaux. Combinez les dans un tableau 2-D de 20 lignes et 2 colonnes.
9. Sauvegardez la matrice obtenue sur le disque.

Thématique 2 (Matplotlib)

1. Créez un tableau NumPy 1-D de la même taille que **heights**, contenant les entiers consécutifs à partir de 1.
2. Affichez le nuage de points (ronds verts) correspondant aux tailles saisies dans la thématique précédente. Soignez la présentation : ajoutez titre, labels d'axes et légende.
3. Affichez une ligne horizontale (en pointillés bleus) représentant la moyenne des tailles, sur la même figure.
4. Enregistrez cette première figure au format *.png sur le disque.
5. Modifiez la figure pour que les graduations de l'axe des ordonnées aille de 1m à 2.2m par pas de 5 centimètres.
6. Affichez un histogramme des tailles, une boite à moustache et une ligne. Sauvegardez chacune des figures au format *.png sur le disque
7. Affichez cote à cote les tailles et les écarts à la moyenne. Soignez la présentation : ajoutez titre, labels d'axes et légende pour chacune des sous-figures.
8. Calculez la corrélation de Pearson entre chaque variable (deux à deux) et affichez la. Les variables sont-elles très corrélées ?

Thématique 3 (Pandas)

1. Chargez le jeu de données iris.csv [1]
2. Affichez les 10 premières lignes du jeu de données.
3. Affichez un sample de 10 lignes, c'est-à-dire 10 lignes prises au hasard.
4. Vérifiez que les données ne contiennent pas de donnée manquante.
5. Affichez les noms des colonnes du jeu de données, sa taille, sa dimension et le type des données.
6. On s'intéresse aux lignes 50 à 78. Créez un nouveau dataframe **iris_subset** contenant uniquement ces lignes. Le nouveau dataframe devra contenir les informations du jeu de données initial, c'est-à-dire les noms de colonne, l'index et les labels.
7. On voudrait que **iris_subset** ne contienne que l'index, les tailles de sépales et l'espèce. Modifiez le code en conséquence, puis affichez les 10 premières lignes de **iris_subset**
8. Sauvegardez **iris_subset** au format CSV sur le disque.
9. Affichez uniquement les observations du jeu de données complet où l'espece est setosa.
10. Affichez uniquement les observations où la largeur des pétales est supérieure à 2cm.
11. Comptez le nombre d'observations que l'on a pour chaque espèce.
12. Pour chaque variable de chaque espèce (séparément), calculez la moyenne et la médiane. Quelle espèce a les sépales les plus petites ? Les plus grandes ?
13. Quelle observation a la pétale la plus longue ? Quelle est sa longueur ? Quelle est son espèce ? Même question pour la pétale la moins longue.
14. On veut ajouter une colonne **total_measure** au jeu de données. Cette colonne sera la somme de toutes les mesures de l'observation. Une fois que c'est fait, sauvegardez le jeu de données sur le disque.

Thématique 4 (Compte rendu)

On souhaite réaliser un compte rendu descriptif du jeu de données iris. En particulier, il sera intéressant d'y trouver : s'il y a des données manquantes, le nombre d'observations, sa dimension, la corrélation entre les variables, des mesures moyennes, une analyse de certaines métriques (moyenne, écart-type, etc.) par espèce, etc. Le compte rendu devra être rédigé en français (ou en anglais) et présenter des visualisations intéressantes permettant d'appuyer le texte. Il pourra s'agir de tableaux ou bien de figures. Le compte rendu ne devra contenir aucun code source ou élément trop technique : il devra être compréhensible par un individu averti mais non expert.

Travail à rendre (noté), et à envoyer à l'adresse mail indiquée en séance :

- Votre compte-rendu
- Vos codes sources

Date limite de remise : voir sur le github.

[1]https://gist.githubusercontent.com/netj/8836201/raw/6f9306ad21398ea43cba4f7d537619d0e07d5ae3/iris.csv