université angers

Licence 3 Informatique
Python

TP2 - Utilisation MatPlot et Pandas

Pour l'ensemble des TP vous pourrez utiliser au choix l'EDI PyCharm ou le docker Jupyter notebook. Pour lancer les éditeurs :

- Ouvrir un terminal,
- `pycharm-community` ou `_jupyter_notebook`
- Créer un nouveau projet pour chaque TP et un fichier pour chaque exercice.

ATTENTION : Chaque TP fera l'objet d'un dépôt sur Moodle qui sera noté à chaque séance.
Date limite de dépôt pour le TP2 **jeudi 13 Février 2020, 23h59.**

Une documentation Python est disponible à cette adresse : https://docs.python.org/3/tutorial/
Une documentation NumPy est disponible à cette adresse : https://docs.scipy.org/doc/numpy/reference/
Une documentation MatPlotLib : https://matplotlib.org/3.1.1/api/index.html
Une documentation Pandas : https://pandas.pydata.org/pandas-docs/stable/reference/index.html

Exercice 1 -Utilisation de NumPy

Les nombres aléatoires

1. Générer un numpy array de 10 valeurs aléatoires
2. Générer un numpy array de 100 valeurs entières entre 0 et 100
3. Générer un numpy array de taille 25 avec des nombres tirés aléatoirement suivant une distribution normale centrée réduite.
4. Générer un numpy array de taille 25 avec des nombres tirés aléatoirement suivant une loi uniforme.
5. Générer une matrice aléatoire de dimension 2 et de taille 10×10.
6. Modifier votre programme pour que les nombres aléatoires générés soient toujours les mêmes à chaque exécution.

Les entrées/sorties

1. Sauvegarder une matrice aléatoire de taille 10×10 dans un fichier matrice1.csv avec ' ;' comme séparateur.
2. Charger le fichier matrice1.csv

Exercice 2 - Utilisation de Matplot

1. À partir de deux tableaux [1,4,8,9] et [12,25,34,78] afficher un nuage de points.
2. À partir d'un tableau numpy aléatoire de 1000 entiers afficher un nuage de points.
3. À partir de deux tableaux [1,4,8,9] et [12,25,34,78] tracer une courbe.
4. À partir d'un tableau numpy aléatoire de 100 nombres suivant une distribution normale tracer une courbe.
5. Tracer trois courbes avec x allant de 0 à 2 avec 100 valeurs, représentant une fonction linéaire, carrée et cubique sur la même figure.
6. Ajouter une légende à ces trois courbes signifiant linéaire, quadratique, cubique.
7. Ajouter un titre "Comparaison des fonctions linéaire, quadratique et cubique".
8. Changer les labels par abscisse et ordonnée.
9. Retracer les trois courbes mais dans 3 figures les unes à côté des autres en attribuant les couleurs bleue, orange et verte.
10. Changer la taille de la figure par (20,7).
11. À partir d'un tableau numpy aléatoire uniforme de 100 éléments allant de 0 à 10 tracer un histogramme constitué de 10 bandes verticales.

Exercice 3 -Activité pratique

Partie 1 : Génération d'un histogramme de la répartition du QI

1. Choisir une graine pour la génération de nombres aléatoires.
2. Créer un numpy array avec une distribution normale de 500 nombres aléatoires.
3. Affecter à cette distribution un coefficient de σ + μ avec par défaut $\sigma = 15$ et $\mu = 100$ ($fx = x \times \sigma + \mu$)
4. Générer l'histogramme correspondant avec 50 bandes verticales.
5. Changer l'axe y par "Probabilité de densité"
6. Changer l'axe x par "Quotient intellectuel"
7. Ajouter un quadrillage
8. Calculer la fonction représentant une loi normale à partir de σ et μ. Rappel de la fonction :

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

   Ici x sera remplacé par $bins$ qui correspond aux valeurs des abscisses de chaque bande verticale de l'histogramme
9. Tracer la fonction y en pointillés.
10. Changer le titre par "Histogramme de la répartition du QI : $\mu = 100,\ \sigma = 15$"

Partie 2 : Gestion de données issues d'un fichier au format csv Charger les fichiers redwineNP.csv et whitewineNP.csv dans deux numpy array redwine et whitewine. Les données sont stockées de cette manière :

0	"fixed acidity"
1	"volatile acidity"
2	"citric acid"
3	"residual sugar"
4	"chlorides"
5	"free sulfur dioxide"
6	"total sulfur dioxide"
7	"density"
8	"pH"
9	"sulphates"
10	"alcohol"
11	"quality"

1. À partir du fichier redwineNP.csv, dans un graphique pyplot :
2. Afficher le degré d'alcool en fonction de sa qualité.
3. Changer les labels pour afficher qualité et degré d'alcool.
4. Afficher la moyenne à l'aide de points rouges pour chaque valeur de qualité.
5. Effectuer le même traitement avec le fichier whitewineNP.csv.
6. Sur un même graphique, comparer les moyennes de dégré d'acools en fonction de la qualité des deux tableaux de données.
7. Pour les deux fichiers, afficher la valeur médiane et les quartiles pour chaque valeur de qualité sous la forme d'un box-plot (boite à moustache).

Exercice 4 -Utilisation de pandas

1. Construire un dataframe de quatre colonnes nommées A, B, C et D et trois lignes numérotées i1, i2 et i3 remplies de 1
2. Construire un dataframe de quatre colonnes nommées p1, p2, p3 et p4 et 100 lignes numérotées de i1 à i100 remplies de 1
3. Modifier le dataframe pour que la colonne p1 contiennent tous les nombres à la puissance 1, jusqu'à p4 contenant tous les nombres à la puissance 4.
4. Afficher la forme du dataframe
5. Afficher les 10 premiers éléments du dataframe

6. Afficher les 10 derniers éléments du dataframe
7. Afficher le noms des colonnes
8. Afficher le type de données des colonnes
9. Afficher les informations complémentaires concernant les colonnes
10. Afficher les statistiques pour le dataframe
11. Afficher les données de la colonne p4 uniquement (2 méthodes)
12. Afficher les données des colonnes p1 et p4 uniquement
13. Afficher le nombres d'éléments appartenant à un intervalle de valeurs sur 10 intervalles dans la colonne p4 (Regarder la documentation sur `value_counts()`)
14. Effectuer la moyenne des ligne et la moyenne des colonnes à l'aide d'un lambda calcul

Exercice 5 -Application de pandas sur des données réelles

1. Charger les fichiers redwinePD.csv et whitewinePD.csv
2. Afficher les 5 alcools avec le plus grand taux d'alcool (avec le plus grand taux en premier puis dans l'ordre décroissant)
3. Afficher les alcools dont la qualité est supérieure à 7
4. Calculer le ratio d'alcools dont la qualité est supérieure à 7 pour le vin rouge et le vin blanc
5. Afficher les vins dont la qualité est supérieure à 7 et la quantité d'alcool inférieure à 10 pour le vin rouge et 9 pour le vin blanc en ne conservant que les colonnes quality et alcohol
6. Afficher la proportion de vin en fonction de leur qualité dans un camembert (pie)
7. Afficher un nuage de points représentant la quantité d'alcool en fonction de la qualité
8. Ajouter une nuance de couleur pour le taux de "volatile acidity"
9. Afficher sous la forme d'une boxplot (boîte à moustache) les différents quartiles de la répartition d'alcool en fonction de la qualité
10. Afficher sous la forme d'un histogramme la répartition des vins en fonction de leur taux d'alcool
11. Afficher plusieurs histogrammes montrant la répartition des vins en fonction de leur taux d'alcool et de leur qualité
12. Afficher la quantité de vins pour un intervalle donné d'alcool et une qualité. On utilisera pour cela un tableau croisé après avoir défini un groupe (avec la fonction cut de pandas) contenant des intervalles pour les degrés d'alcool. On fera l'affichage avec 10 intervalles.