TP1 : Analyse exploratoire avec seaborn

Pour ce TP d'initiation à seaborn, nous allons utiliser le jeu de données "wine-quality-white-and-red.csv" extrait de l'article *Modeling wine preferences by data mining from physicochemical properties* de Cortez *et al.* paru dans *Decision Support Systems* en 2009.

Ce jeu de données contient différentes mesures chimiques sur plusieurs vins rouges et blanc.

Nous nous contenterons dans cette séance d'utiliser des fonctions de base.

Documentation :
http ://www.python-simple.com/python-pandas/panda-intro.php
https ://www.kaggle.com/code/jsaguiar/exploratory-analysis-with-seaborn
https ://pandas.pydata.org/docs/reference/frame.html

Travail à réaliser :
Réaliser un notebook qui réalise une analyse exploratoire du jeu de données

1 Visualisation

1. créer une variable mydata à partir de la lecture du fichier csv avec
   pd.read_csv('wine-quality-white-and-red.csv')
2. réaliser une copie de "mydata" en "df"
3. afficher les 5 premières lignes : df.head(5)
4. Afficher une colonne particulière du jeu de données
5. Afficher une valeur particulière
6. Afficher uniquement les vins de type "rouge" puis "blancs"
7. Citer les données catégoriques du dataset
8. Décrire le dataset à partir de la fonction describe
9. Analyser le dataset avec les différentes options de pairplot :
   https ://seaborn.pydata.org/generated/seaborn.pairplot.html
10. Afficher le pH en fonction des chlorides paramétrée en fonction du niveau d'alcool (hue='alcohol') en utilisant la fonction scatterplot
11. Afficher la quantité de vin rouge et de vins blanc en fonction de sa qualité en utilisant countplot
12. Utiliser pairplot et commenter le résultat.
13. Utiliser barplot et commnenter le résultat
14. Remplacer le type "red" par 1 et le type "blanc par "zero"

2 Valeurs manquantes

1. Supprimer 10 valeurs du fichier csv et les remplacer par NaN
2. Afficher le headmap des valeurs manquantes (sns.heatmap(df.isna().transpose(), cbar=False, ax=ax))
3. supprimer les lignes où il y a une valeur manquante.
4. Utiliser .drp
5. Remplacer les valeurs manquantes par interpolation

3 Etude statistique

1. Calculer les moyennes et écart-type des pH des vins rouges et blancs
2. Afficher le diagramme pour le pH des vins rouges
3. Afficher le diagramme pour le pH des vins blancs
4. Afficher l'histogramme de toutes les variables
5. Comparer la répartition des données en utilisant les boîtes à moustache
6. Comment se lisent ces graphiques ? Que nous disent-ils sur les répartitions des variables alcohol et pH
7. Calculer les quartiles et la médiane du taux d'alcool dans les vins rouges et celui dans les vins blancs. Commentez les résultats obtenus
8. Quel est le poids du groupe des vins rouges ? et celui du groupe des vins blancs ?
9. Calculez la moyenne et l'écart type de chaque groupe.
10. Utiliser la fonction ".describe" et comparer vos résultats avec ceux des 2 questions précédentes
11. Afficher la matrice de corrélation : commenter le résultat

4 Prétraitement élémentaire

1. Réaliser une normalisation des données via StandardScaler
2. Réaliser une normalisation des données via MinMaxScaler
3. Analyser l'impact de la normalisation