TP2 : Bibliothèque Pandas : Analyse de données complexes

Enoncé 1 : Data Cleaning

On se propose de travailler sur un ensemble de données contenant 18173 lignes relatives aux activités de ventes dans une société.

L'entête des données étant la suivante :

- Series_reference
- Period
- Data_value
- STATUS
- UNITS
- Subject
- Group
- Series_title_1

Après avoir téléchargé le contenu du fichier CSV dans un DataFrame, on vous demande d'intégrer les traitements suivants sur les données comme suit :

- Changer le nom des deux premiers champs respectivement par Ref et Période
- Supprimer la dernière colonne qui n'est pas utile pour notre analyse
- Nettoyer les données en supprimant les valeurs NaN
- Effectuer un copie du DataFrame (avec ou sans élimination de redondances) et faire une copie vers un fichier nommé Formatte.CSV.
- Sur la base de cette nouvelle copie supprimer à la fois les colonnes UNIT et Subjets
- Supprimer les lignes dont la valeur du champ STATUS est égale à ''FINAL''

Enoncé 2 :

On souhaite récupérer les données donnees_enquete_2003_television.csv qui reporte des données relatives à l'audience télévisée.

Les champs décrits dans cette source sont les suivants :

- POIDSLOG : Pondération individuelle relative

- POIDSF : Variable de pondération individuelle
- cLT1FREQ : Nombre d'heures en moyenne passées à regarder la télévision
- cLT2FREQ : Unité de temps utilisée pour compter le nombre d'heures passées à regarder la télévision, cette unité est représentée par les quatre valeurs suivantes :
    - 0 : non concerné
    - 1 : jour
    - 2 : semaine
    - 3 : mois

On vous demande de :

- Supprimer les colonnes vides
- Obtenir les valeurs distinctes pour la colonne cLT2FREQ
- Modifier la matrice pour enlever les lignes pour lesquelles l'unité de temps (cLT2FREQ) n'est pas renseignée ou égale à zéro.
- Sauver le résultat au format Excel.

Enoncé 3 : Interrogation des données

On se propose d'intégrer des analyses avancées sur un ensemble données en utilisant les opérations de sélection, tri, groupement etc.

Les données source de notre étude se trouvent dans le fichier : ''Automobiles.csv''.

La source de données contient des colonnes non significatives à supprimer pour ne garder que les données suivantes : "index","company","body-style","wheel-base", "length", "engine-type", "num-of-cylinders", "horsepower", "average-mileage", "price".

Question 1 : Commencer par vérifier le type de données et afficher les 5 premières lignes du jeu de données.

Question 2 : Ensuite, nettoyer les données en éliminant les valeurs NaN et 0 et reporter une nouvelle version dans le même fichier.

Question 3 : Afficher le nom de la compagnie, le nombre de chevaux ainsi que le prix de la voiture la plus chère

Question 4 : Afficher le détail de toutes les voitures de marque Toyota

Question 5 : Afficher le nombre de voitures par compagnie

Question 6 : afficher pour chaque compagnie le prix de la voiture le plus élevé

Question 7 : afficher la moyenne de kilométrage par compagnie

Question 8 : trier toutes les voitures par prix

Question 9 : Interprétez le code suivant :

```
GermanCars = {'Company': ['Ford', 'Mercedes', 'BMV', 'Audi'], 'Price': [23845, 171995, 135925 , 71400]}
carsDf1 = pd.DataFrame.from_dict(GermanCars)

japaneseCars = {'Company': ['Toyota', 'Honda', 'Nissan', 'Mitsubishi '], 'Price': [29995, 23600, 61500 , 58900]}
carsDf2 = pd.DataFrame.from_dict(japaneseCars)

carsDf = pd.concat([carsDf1, carsDf2], keys=["Germany", "Japan"])
```