Traitement des données en Python

Université d'Angers – M1 Data Science

Quelques manipulations de `pandas`

Exercice 1. Applications immédiates d'un dataframe.

1) Écrire une fonction `simu(n)` qui simule un dataframe de taille $n \times 5$ contenant : `index` $= i$, `nom` = 'enregistrement i', `val1` = valeur aléatoire entre 0 et 1, `val2` = valeur aléatoire entre 0 et 1 et `val3` = valeur aléatoire entre 0 et 1.
2) Appeler la fonction (avec par exemple $n = 10$) et afficher le dataframe obtenu.
3) Ajouter une colonne `moy` qui calcule la moyenne des 3 valeurs sur chaque ligne.
4) Supprimer les colonnes `val1`, `val2` et `val3`.
5) Créer et afficher un nouveau dataframe identique à l'ancien mais qui ne contient que les lignes dont la moyenne est ≥ 0.5.

Exercice 2. Récupérer le jeu de données `titanic.csv`.

1) Charger les données (colonnes 0, 1, 2, 4, 5) avec la colonne 0 pour index.
2) Afficher le nombre de passagers.
3) Combien y a-t-il d'hommes et de femmes ? Combien y a-t-il de survivants ?
4) Quelle est la proportion d'hommes et de femmes ayant survécu ? Par classe (1, 2 ou 3) ?
5) Quelle est la proportion de survivants ayant moins de 18 ans et plus de 18 ans ?

Exercice 3. Récupérer le jeu de données `workers.data`.

1) Charger les données (colonnes 0, 9, 12, 13, 14) et nommer les colonnes (ex : `Age`, `Sexe`, `Heures/Sem`, `Pays` et `Salaire`).
2) Extraire un échantillon de 10000 lignes du jeu de données, choisies aléatoirement, et trier cet échantillon par index croissant avant de le réindexer (de 0 à 9999).
3) Reprendre le jeu de données initial et appeler `Num` l'index.
4) Récupérer la liste des pays en supprimant les doublons, et les trier par ordre alphabétique.
5) Construire le tableau croisé des sexes par rapport aux salaires.
6) Calculer la moyenne d'heures de travail par semaine des indiens dont le salaire est > 50K.
7) Donner la liste des 5 pays dans lesquels la moyenne d'heures de travail par semaine est la plus élevée (par ordre décroissant).

Exercice 4. Récupérer le jeu de données `etatcivil.csv`.

1) Charger les données et nommer les colonnes (ex : `Année`, `Nom`, `Sexe` et `Naissances`).
2) Sélectionner tous les enregistrements dont le prénom commence par "Fran". Établir la liste alphabétique de ces prénoms.

3) Quels sont les 5 prénoms commençant par 'Fran' les plus donnés pendant cette période ? Tracer sur un même graphique l'évolution par années du nombre d'apparitions de ces 5 prénoms.

4) Au fil des années, le prénom Camille a été un prénom essentiellement masculin ou essentiellement féminin. Tracer les courbes représentant, par années, le pourcentage de filles et de garçons parmi les Camille.

Exercice 5. Récupérer le jeu de données `isd-history.csv`.

1) Charger les données (on pourra conserver les noms de colonnes indiquées dans le fichier) avec `USAF` comme index.

2) Combien y a-t-il d'enregistrements ?

3) Compter le nombre de stations dans chaque hémisphère (latitude positive ou négative) à l'aide d'une coupure. Indiquer également le nombre d'enregistrements dont la latitude est manquante.

4) Convertir les colonnes `BEGIN` et `END` au format date `YYYY-MM-DD`.

5) Déterminer les 10 stations qui ont eu la période d'activité la plus grande. On pourra créer une nouvelle colonne avec la période d'activité exprimée en jours.

6) Déterminer le nombre de pays ayant des stations ainsi que celui ayant le plus de stations.