Activité guidée : Découverte de Pandas

Pandas - Matplotlib - SQL

Mise en situation :

Félicitations, vous venez de décrocher votre premier job au sein de la FAO (**Organisation des Nations Unies pour l'alimentation et l'agriculture).**

Votre première mission est de réaliser une étude sur la démographie et la nutrition.

Étape n°0 : se former sur pandas et préparer son environnement

Commencez par revoir ou apprendre les concepts principaux de Pandas avant de vous lancer dans le projet, rendez-vous ici.

Pour ce projet, je vous conseille de travailler sur votre environnement conda dédié à la data science et d'utiliser un jupyter notebook pour répondre aux questions.

Étape n°1 : Récupérer les jeux de données

Rendez-vous sur le site de la FAO dans la section FAOSTAT puis dans la catégorie bilans alimentaires.

Votre objectif est d'extraire 3 jeux de données (fichiers csv) avec les informations suivantes :

- les produits d'origine animale
- les produits d'origine végétale
- la population de chaque pays

Sélectionnez les années **2018** et **2019.**

Pour les *data sets* relatifs aux produits, sélectionnez les éléments suivants (4):

- disponibilité alimentaire en kg
- disponibilité alimentaire en kcal
- disponibilité alimentaire en protéines
- disponibilité alimentaire en matière grasse

Vous aurez donc 3 fichiers csv. Importez les données avec **Pandas** afin de créer 3 DataFrames : **df_anim**, **df_veg** et **df_pop**.

Avant de vous lancez dans le code, prenez du temps pour comprendre les données et la mission de la FAO. Pour cela : explorez la section définitions et standards

Étape n°2 : Nettoyage et préparation des données

Explorez votre jeu de données grâce à des méthodes de la librairie **Pandas**. Les méthodes utilisées doivent vous permettre de répondre aux questions suivantes :

1. Nettoyer les titres de colonnes :
    - Supprimer les espaces au début et à la fin des titres (s'il y en a)
    - Remplacer les espaces par des underscores (ceux se situant entre les mots)
    - Tout mettre en minuscule
2. Quelle sont les dimensions des jeux de données ?
3. A quoi ressemblent les 5 premières lignes de mes jeux de données ?
4. Pour les datasets **df_anim** et **df_veg**, ajoutez une colonne 'type' qui prendra respectivement une valeur 'animal' et 'vegetal'. Une fois cette étape effectuée, regroupez les deux jeux de données en 1 et appelez ce DataFrame **product**. Attention à bien comprendre la structure des données pour utiliser la bonne méthode.
5. Transformez **df_pop** afin de ne garder que le code du pays, le pays, l'année et la population. Renommer la colonne 'value' en 'pop_1000_hab'.
6. Transformez **products** afin de ne garder que les colonnes *area_code_(fao)*, *area*, *element*, *item*, *year*, *type*, *unit*, *value.*
7. Fusionnez **df_pop** avec **products** et nommez ce DataFrame **df**. Afin de fusionner ces jeux de données vous devez identifier les clés primaires. Renommer les colonnes comme sur le capture d'écran ci-dessous.

   Note : ajouter la population à la base de données créera de la redondance dans la bdd mais simplifiera les calculs pour les questions suivantes. Ne pas supprimer df_pop.

   A ce stade votre jeu de données doit ressembler à cela :

	area_code_(fao)	area	year	pop_1000_hab	element	item	type	value
0	2	Afghanistan	2018	37172.0	Food supply quantity (kg/capita/yr)	Wheat and products	vegetal	160.12
1	2	Afghanistan	2018	37172.0	Food supply (kcal/capita/day)	Wheat and products	vegetal	1372.00
2	2	Afghanistan	2018	37172.0	Protein supply quantity (g/capita/day)	Wheat and products	vegetal	37.00
3	2	Afghanistan	2018	37172.0	Fat supply quantity (g/capita/day)	Wheat and products	vegetal	4.59
4	2	Afghanistan	2018	37172.0	Food supply quantity (kg/capita/yr)	Rice and products	vegetal	19.78

   <u>Maintenant que nous avons un DF unique, il sera plus facile de l'explorer :</u>

8. Quelles sont les types de données de chaque colonne ?
9. Combien y-a t'il de valeurs manquantes par variable ?
10. Est-ce qu'il y a des valeurs aberrantes ? (population négative, etc.) Utilisez un récapitulatif statistique pour répondre à cette question.
11. Affichez les valeurs uniques de la colonne **area**
12. Gardez uniquement les informations relatives aux pays (supprimez les zones géographiques ou économiques) Note : en fonction de votre méthode d'importation de données cette étape est facultative.

13. Modifiez votre jeu de données afin que les informations soient indexées par **area_code**, **area**, **year**, **pop_1000_hab, type et item**. Les valeurs de la colonne **element** doivent être séparées dans des colonnes différentes. Recherchez sur internet la différence entre les formats *long* et les formats *wide*. Pour réussir cette étape creuser la méthode pivot_table. Il est préférable d'appliquer la méthode reset_index() après avoir utilisé la méthode précédente.
14. Faire du nettoyage dans le nom des colonnes

    À ce stade votre jeu de données doit ressembler à cela :

	area_country	country	year	item	pop_1000_hab	fat_supply_quantity_(g/capita/day)	food_supply_(kcal/capita/day)	food_supply_quantity_(kg/capita/yr)	protein_supply_qu
0	1	Armenia	2018	Apples and products	2952	0.12	20.0	14.49	
1	1	Armenia	2018	Aquatic Animals, Others	2952	0.00	0.0	0.00	
2	1	Armenia	2018	Aquatic Plants	2952	0.00	0.0	0.00	
3	1	Armenia	2018	Bananas	2952	0.06	13.0	7.61	
4	1	Armenia	2018	Barley and products	2952	0.03	6.0	0.72	
...	...	...	...	...	...	...	...	...	
30140	276	Sudan	2019	Tomatoes and products	42813	0.10	9.0	15.36	
30141	276	Sudan	2019	Vegetables, other	42813	0.14	20.0	27.51	
30142	276	Sudan	2019	Wheat and products	42813	1.67	535.0	61.56	
30143	276	Sudan	2019	Wine	42813	0.00	0.0	0.00	
30144	276	Sudan	2019	Yams	42813	0.02	12.0	4.23	

30145 rows × 9 columns

    Notez bien la dimension du jeu de données. Si tout s'est bien déroulé vous devez avoir le même resultat, sinon revoyez les étapes précédentes.

15. Créez des masques afin d'afficher un DataFrame qui ne contient que l'année 2018
16. Nous allons ajouter une nouvelle colonne à notre jeu de données : la zone géographique. J'ai récupéré ces informations pour vous sur le site de la FAO. Effectuer un merge entre ce jeu de données et le vôtre.
17. *Bonus : Créer ce jeu de données par vous même sur le site de la FAO.*

Étape n°3: Exploration

Lorsqu'aucune précision n'est donnée, veuillez répondre en utilisant l'année la plus récente :

1. Quelle est la médiane de la variable **fat_supply_quantity_(g/capita/day)** ? Q1 ? Q3 ? La moyenne ? L'écart type ? (il existe une méthode pour visualiser toutes ces informations en même temps). Interprétez ces mesures statistiques dans une phrase.
2. Visualisez la distribution des données numériques à l'aide d'un histogramme. Utilisez une boucle si nécessaire. Si certaines variables contiennent des valeurs extrêmes vous pouvez effectuer une transformation sur vos données. (En logarithmes par exemple)

3. Quelle est la population de l'Ukraine en 2018 ? L'output doit être un int (pas un DataFrame)
4. Quels sont les 10 pays les plus peuplés ?
5. Quelle est la population mondiale en 2018 ? En 2019 ? Est-ce que ce chiffre correspond à la réalité ? Menez votre enquête et faites les corrections nécessaires en cas de problème. Contrôlez vos données grâce à ce site.
6. Pour quels pays dispose-t-on du moins d'informations (nombre de valeurs manquantes) ? Donnez-en 5.
7. Créez une nouvelle colonne **taux_croissance_pop_18_19** avec le taux de variation de la population entre 2018 et 2019 dans chaque pays. Affichez les 5 pays avec le taux de croissance démographique le plus élevé.
8. Quel est le taux de croissance moyen en fonction de la zone géographique ?
9. Calculez la disponibilité de nourriture totale par pays et par année, en kcal et kg de protéines. Attention aux unités de mesure !
10. Calculez le ratio énergie/poids de chaque produit et pays. Vous devriez vous apercevoir qu'étonnement, ces informations varient en fonction du pays. Pour pallier ce problème, calculez la moyenne de ce ratio pour chaque aliment. Attention à bien gérer les valeurs égales à 0. Vérifiez la cohérence de votre calcul en comparant le résultat avec l'apport calorique d'un œuf.
11. À l'instar de la question précédente, calculez le pourcentage de protéine de chaque aliment. Vérifiez votre résultat en le comparant avec l'apport en protéines d'un œuf.
12. Quels sont les 10 aliments les plus caloriques ?
13. Quels sont les 10 aliments les plus riches en protéines ?
14. Créez une boîte à moustache de la quantité de nourriture par habitant en kcal par pays. Créez sur un même graphique un boxplot par zone géographique. Mettez un titre, changez les étiquettes des axes et changez la couleur en fonction de la zone géographique. Effectuez ce graphique en utilisant la librairie de visualisation **matplotlib** puis installez la librairie **seaborn** et refaites-le. Commentez les différences d'utilisation entre ces deux librairies.
15. Question bonus : Avec la disponibilité alimentaire de produits végétaux combien d'être humains pourrait-on nourrir ?