Exercice partie 2 - Analysez des données de systèmes éducatifs

Conseils mentor

L'étudiant peut potentiellement réaliser cet exercice en quelques heures, sans même une session de mentorat.

Cependant, s'il a des difficultés pendant l'exercice et que vous avez une session de mentorat, demandez-lui de vous parler de :

1. son avancement ;
2. ses difficultés avec la syntaxe du langage Python et la compréhension des concepts de base de programmation, dans le cas où il n'aurait jamais codé par le passé.

Vous pourrez ainsi identifier ses problèmes, ses lacunes, ses processus ou méthodologies erronés.

- **Compréhension des attendus**

Assurez-vous que l'étudiant puisse répondre à ces questions :

- Pourquoi appliquons-nous toutes ces opérations de filtrage ?
  - Si vous le sentez en difficulté pour vous répondre, n'hésitez pas à l'orienter sur la notion de "trop grand nombre d'indicateurs" pour arriver à un périmètre exploitable manuellement.
- Pourquoi le hardcoding fait-il perdre du temps sur le long terme ?
- Pourquoi utilisons-nous une agrégation pour ne plus se situer à la maille indicateur + année + pays ?

- **Lacunes récurrentes**

Si vous sentez que votre étudiant n'est pas à l'aise avec l'utilisation de Pandas, n'hésitez pas à lui conseiller de revoir la partie 2 du cours "Nettoyez et analysez votre jeu de données" avant d'aller plus loin et de vous en faire un résumé oral.

- **Méthodologie**

Les analyses exploratoires sont "floues" et subjectives par nature. Si vous sentez que l'étudiant ne comprend pas comment les différentes étapes aboutissent à une réponse à la demande de Mark, parlez-lui de votre expérience avec des projets d'analyse exploratoire et comment ils ont permis d'accomplir un objectif métier.

- **Points de vigilance**

Concernant les étudiants qui n'ont jamais codé, il est fréquent qu'ils aillent trop loin dans l'apprentissage des généralités de Python, alors que l'objectif est de se focaliser sur l'usage de Python pour l'analyse de données. Assurez-vous que l'étudiant se concentre sur cet aspect-là de l'apprentissage du langage.

Voici le document qui vous aidera à faire un bilan avec l'étudiant :

- Sa fiche d'autoévaluation
- Elle permet de faire le point sur ce que l'étudiant a appris. Il la complète seul et en discute avec vous.

Prêt à résoudre l'exercice ?

Dans cet exercice, vous allez apprendre à réduire le périmètre de la donnée (nombre de lignes et colonnes).

Après avoir analysé en surface les différents fichiers, vous avez remarqué :

- que le fichier Data est plus important que les autres ;
- que vous avez un nombre d'indicateurs beaucoup trop élevé pour vous permettre une analyse manuelle de chacun.

Vous devez **réduire** le périmètre de la donnée pour **faciliter le choix des indicateurs** et répondre à la demande de Mark.

Cet exercice est entièrement guidé, suivez les étapes ci-dessous.

Étapes

Étape 1 - Réduisez le périmètre en utilisant une approche métier

Prérequis

- Avoir obtenu un dataframe par fichier après avoir suivi les étapes de l'exercice précédent.

Résultat attendu

- Du code et du markdown dans le même Jupyter Notebook, permettant de reproduire les réponses aux instructions ci-dessous.

Instructions

- Parmi les jeux de données centrés sur les indicateurs, identifiez la colonne qui décrit la catégorie métier à laquelle appartient chaque indicateur.
- Gardez les catégories qui font sens par rapport à la demande de Mark et l'objectif du projet et supprimez les autres.
- Calculez le nombre d'indicateurs restants.
- Filtrez l'ensemble des jeux de données pour ne garder que les indicateurs sélectionnés.
- Dans le fichier Data, interprétez les colonnes représentant les années.
  - Pourquoi avons-nous des valeurs d'indicateur pour des années futures ?
  - Sur la base de votre compréhension de la problématique métier, filtrez les années en conséquence.

Recommandations

Relisez bien le projet pour bien comprendre le but final de l'analyse : projet d'expansion à l'international.

- La méthode value_counts() de Pandas est particulièrement utile pour comprendre le contenu d'une colonne catégorielle.
- Réutilisez la même approche de filtrage que dans l'exercice précédent.
  - Le filtrage en utilisant la méthode isin() de Pandas vous permettra également de filtrer les colonnes en précisant une liste d'années à supprimer.
  - Construisez cette liste d'années avec la méthode np.arange() du package numpy

- Vu qu'il y a plusieurs années disponibles pour chaque indicateur, réfléchissez pour savoir quelles années sont pertinentes ou non :
  - par rapport à ce que l'on cherche ;
  - par rapport au taux de remplissage/nombre de valeurs manquantes.

Points de vigilance

- Dans un projet data, il y a rarement une seule et unique bonne réponse.
- Vous allez forcément faire appel à votre bon sens et votre compréhension subjective de la demande de Mark pour sélectionner certaines catégories d'indicateur et pas d'autres. Vous allez décider différemment par rapport aux cas limites.

Tant que vos hypothèses sont raisonnables, bien documentées et faciles à modifier, votre travail sera considéré correct. N'ayez pas peur de vous tromper, vous êtes là pour apprendre et votre mentor pourra vous orienter si besoin.

Ressources

- La partie 2 "Nettoyez un jeu de données" du cours.

Étape 2 - Réduisez le périmètre en utilisant une approche data

Résultats attendus

- Du code et du markdown dans le même Jupyter Notebook, permettant de reproduire les réponses aux instructions ci-dessous.
- Une variable contenant une liste d'une quinzaine maximum d'indicateurs sélectionnés.

Instructions

- Pour chaque année, calculez la proportion d'indicateurs avec des valeurs renseignées (c'est-à-dire, non manquantes).
- Pour chaque indicateur, calculez la proportion d'années avec des valeurs renseignées.
- Sur la base des opérations précédentes, réduisez le nombre d'années et d'indicateurs pour garder celles et ceux qui sont les plus riches en données (et donc les plus utilisables pour la suite).
  - Identifiez les indicateurs particulièrement riches en données en calculant un dataframe contenant par indicateur et par année, le nombre de pays avec une valeur renseignée.
  - Triez ce dataframe par ordre décroissant de nombre de pays avec une valeur renseignée, afin d'obtenir les indicateurs les plus riches en données dans l'ensemble.
  - Aidez-vous de ce dataframe pour sélectionner parmi les indicateurs les plus riches, une quinzaine qui font sens par rapport à la problématique métier.

Recommandations

- Servez-vous de la fonction groupby() que vous avez vue en cours pour faciliter le calcul des différentes proportions.

Soignez le nommage de vos variables, à ce stade vous avez déjà manipulé plus d'une dizaine de dataframes. Si vous les appelez tous df_1, df_2, df_14... vous risquez de vous perdre, voire de ne plus comprendre votre propre code quand vous allez le relire demain ou la semaine prochaine ! N'hésitez pas à aligner le nom de vos variables avec la donnée qu'elles contiennent.
Par exemple, le résultat de la première instruction pourrait s'appeler : proportion_indicateurs_par_annee (évitez les caractères spéciaux et les majuscules dans les noms de variables).

Points de vigilance

- Même les indicateurs les plus riches en données peuvent probablement présenter un nombre assez important de valeurs manquantes.
- Trouvez le juste équilibre entre quantité d'indicateurs, pertinence métier des indicateurs et complétude en données de ceux-ci.

Étape 3 - Consolidez vos résultats dans un dataframe (pays, indicateurs)

Résultats attendus

- Du code et du markdown dans le même jupyter notebook, permettant de reproduire les réponses aux instructions ci-dessous.
- Un dataframe avec la structure décrite dans la deuxième instruction.

Instructions

- Filtrez votre dataframe Data pour ne garder que les indicateurs, pays et années que vous avez jugé pertinents sur la base de vos analyses précédentes.
- Agrégez ce dataframe pour en construire un nouveau : chaque ligne doit correspondre à un pays et chaque colonne doit correspondre à un indicateur.

Recommendations

- Le dataframe Data que vous avez est à la maille (indicateur, année, pays), vu le format du dataframe demandé, vous devez agréger les années pour chaque pays et indicateur. La méthode la plus simple ici est d'utiliser pivot_table() et non groupby() (également possible, mais plus complexe).
- Réfléchissez à comment vous voulez résumer vos années en une seule statistique agrégée. Plusieurs choix sont possibles : moyenne, médiane, moyenne pondérée etc.

Points de vigilance

- Si vous avez bien mené votre analyse jusqu'à présent, vous ne devriez pas avoir beaucoup de combinaisons (pays, indicateurs) avec trop peu d'années renseignées.
- Le cas inverse poserait problème, car calculer des moyennes sur une année ou deux serait beaucoup moins fiable que d'en calculer sur quatre ou cinq. Demandez conseil à votre mentor si c'est le cas.

Étape 4 - Vérifiez votre travail et faites le point avec votre mentor

Pour vérifier que vous n'avez rien oublié dans la réalisation de votre exercice, téléchargez et complétez la **partie 2** de la fiche d'autoévaluation.

Parlez-en avec votre mentor durant votre dernière session de mentorat.

Avez-vous une suggestion pour nous ?