**Analysez des données de systèmes éducatifs**

**Enoncé :**

Vous êtes Data Scientist dans une start-up de la EdTech, nommée academy, qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

Mark, votre manager, vous a convié à une réunion pour vous présenter le projet d’expansion à l’international de l’entreprise. Il vous confie une première mission d’analyse exploratoire, pour déterminer si les données sur l’éducation de la banque mondiale permettent d’informer le projet d’expansion.

Mark aimerait explorer les pays avec un fort potentiel de clients pour les services de academy, et comment ce potentiel pourrait évoluer.

Votre mission :

Mark vous a donc demandé de réaliser une analyse pré-exploratoire de ce jeu de données.

Il vous a transmis ce courriel à la suite de la réunion :

*Hello,*

*Les données sont disponibles sur "ce site de la Banque mondiale".*

*Je te laisse regarder la page d'accueil qui décrit le jeu de données.*

*En résumé, l’organisme “EdStats All Indicator Query” de la Banque mondiale répertorie 4 000 indicateurs internationaux décrivant l’accès à l’éducation, l’obtention de diplômes et des informations relatives aux professeurs, aux dépenses liées à l’éducation... Tu trouveras plus d'info sur ce site.*

*Pour la préanalyse, pourrais-tu :*

*> Valider la qualité de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?)*

*> Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?)*

*> Sélectionner les informations qui semblent pertinentes pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l’entreprise ?)*

*> Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type par pays et par continent ou bloc géographique)*

*Ton travail va nous permettre de déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays. On va partager ton analyse avec le board, alors merci de soigner la présentation et de l'illustrer avec des graphiques pertinents et lisibles !*

*Merci !*

**Livrables :**

* Un Jupyter Notebook comportant les analyses pré-exploratoires réalisées.
* Un support de présentation pour la soutenance (15 mn).

**Soutenance :**

Pendant la soutenance, l’évaluateur jouera le rôle de Mark. Vous lui présenterez l’ensemble de votre travail.

* Présentation (20 minutes)

Rappel de la problématique et présentation du jeu de données (5 minutes)

Présentation de l’analyse pré-exploratoire du jeu de données et vos conclusions sur la pertinence de l’usage du jeu de données pour répondre aux questions stratégiques que se pose l’entreprise (15 minutes)

* Discussion (5 minutes)

L’évaluateur vous challengera sur vos choix.

* Débriefing (5 minutes)

À la fin de la soutenance, vous pourrez débriefer ensemble.

Votre présentation devrait durer 15 minutes (+/- 5 minutes).

Puisque le respect des durées des présentations est important en milieu professionnel, les présentations en dessous de 10 minutes ou au-dessus de 20 minutes peuvent être refusées.

**Référentiel d'évaluation :**

Compétences évaluées :

> Mettre en place un environnement Python

> Maîtriser les opérations fondamentales du langage Python pour la Data Science

> Manipuler des données avec des librairies Python spécialisées

> Effectuer une analyse univariée et des représentations graphiques avec une librairie

> Utiliser un notebook Jupyter pour faciliter la rédaction du code et la collaboration

Mettre en place un environnement Python :

CE1 Vous avez installé le logiciel Python sur votre ordinateur.

CE2 Vous avez installé les librairies Python nécessaires pour réaliser le projet.

CE3 Vous avez installé un nouveau package depuis Jupyter en utilisant un gestionnaire de package (ex : pip).

CE4 Vous avez vérifié les numéros de version des packages installés, afin de s'assurer d'avoir des versions récentes.

CE5 Vous avez créé un environnement virtuel pour assurer l’isolement du projet et la gestion des dépendances.

Maîtriser les opérations fondamentales du langage Python pour la Data Science:

CE1 Vous avez codé au moins une fois une répétition d’une action à l'aide d'une boucle "for".

CE2 Vous avez utilisé au moins une fois un opérateur de comparaison avec une variable.

CE3 Vous avez utilisé au moins une fois des conditions avec un opérateur de comparaison.

CE4 Vous avez créé au moins un objet dataframe.

Manipuler des données avec des librairies Python spécialisées :

CE1 Vous avez chargé en mémoire un fichier plat.

CE2 Vous avez décrit un jeu de données avec au moins les informations suivantes :

Nombre de colonnes, nombres de lignes, types des données pour chaque colonne, taux de valeurs manquantes par colonne, moyenne/médiane/écart-type par colonne.

CE3 Vous avez effectué chacune de ces opérations : sélection de colonne, sélection de ligne, filtrage et opération d'agrégation sur un dataframe.

Effectuer une analyse univariée et des représentations graphiques à l'aide d'une librairie Python adaptée :

CE1 Vous avez calculé les moyennes, médianes et quantiles des distributions.

Vous avez tracé au moins une représentation graphique d’une distribution statistique.

CE2 Les éventuelles valeurs aberrantes ont été identifiées, selon une approche statistique (méthode interquartile) et une approche métier (par exemple min/max des valeurs réelles).

CE3 Vous avez réalisé plusieurs représentations graphiques dans l'environnement de développement via une librairie Python spécialisée. Les graphiques sont lisibles (titres alignés, légende présente, noms des abscisses et des ordonnées précisés).

Utiliser un notebook Jupyter pour faciliter la rédaction du code et la collaboration :

CE1 Vous avez utilisé un notebook Jupyter.

CE2 Vous avez utilisé des cellules Markdown pour structurer et commenter le notebook.

Vous avez installé une extension pour afficher automatiquement la “table of content” du notebook.