

# Auto-évaluation

Renforcez vos fondamentaux de Python pour la Data Science en analysant les données de systèmes éducatifs

Exercice en 3 parties - Analysez des données de systèmes éducatifs



Un dernier doute ?

Pour vérifier la qualité de votre travail :

- cochez les cases ci-dessous : elles indiquent que vous avez bien pris en compte chaque indicateur de réussite ;
- complétez votre fiche au fur et à mesure de votre progression dans votre exercice.
- renseignez, si besoin, la colonne "Notes" avec des commentaires sur vos livrables / vos étapes. Ils seront des points de discussion avec votre mentor pendant votre session de bilan / soutenance.

Quand toutes les cases de ce document seront cochées, vous pourrez déposer vos livrables sur la plateforme.

Bonne réussite !

| Compétences  | Indicateurs de réussite de l'activité   | Notes |
|--|---|-------|
| <b>Exercice - Partie 1</b>   |   |       |
| Appliquer des analyses statistiques descriptives et / ou naviguer visuellement au sein des données afin de | <input type="checkbox"/> J'ai vérifié que l'installation de JupyterLab (ou un équivalent) était fonctionnelle.<br><br><input type="checkbox"/> J'ai importé Pandas, Matplotlib et Seaborn sans erreur au sein du Notebook.<br><input type="checkbox"/> Je suis capable d'expliquer ce que représente chaque ligne dans chaque jeu de données.<br><br><input type="checkbox"/> J'ai calculé la taille des jeux de données avec Pandas. |       |
|  |   |       |

|  |   |  |
|--|---|--|
| détecter des anomalies.  | <input type="checkbox"/> J'ai décrit le contenu des colonnes de chaque jeu de données en utilisant les méthodes fournies par Pandas.  |  |
| Supprimer / corriger les anomalies manuellement et à l'aide d'outils / logiciels de nettoyage des données adaptés. | <input type="checkbox"/> J'ai calculé le nombre de doublons avec Pandas et les ai supprimés.<br><input type="checkbox"/> J'ai quantifié les valeurs manquantes par colonne dans les jeux de données.<br><input type="checkbox"/> Mon traitement des valeurs manquantes est en cohérence avec la problématique métier.<br><input type="checkbox"/> J'ai utilisé les filtrages à base de conditions de Pandas : <ul style="list-style-type: none"> <li><input type="checkbox"/> pour supprimer les faux pays ;</li> <li><input type="checkbox"/> pour réduire le nombre d'indicateurs via une colonne contenant des catégories d'indicateurs.</li> </ul> <input type="checkbox"/> Ma matrice de corrélation est cohérente : <ul style="list-style-type: none"> <li><input type="checkbox"/> elle a été calculée avec Pandas ;</li> <li><input type="checkbox"/> visualisée avec Seaborn ;</li> <li><input type="checkbox"/> utilisée pour filtrer les indicateurs redondants.</li> </ul> <input type="checkbox"/> Je suis capable d'expliquer le sens des coefficients de corrélation de Pearson et de Spearman.<br><input type="checkbox"/> Je suis satisfait de l'utilisation des markdown. |  |
| <div>Exercice - Partie 2</div>   |   |  |
| Appliquer des analyses statistiques descriptives et /  | <input type="checkbox"/> J'ai utilisé la méthode group_by() de Pandas pour évaluer la qualité des indicateurs et des années (en termes de valeurs renseignées).   |  |

|   |  |  |
|---|--|--|
| ou naviguer visuellement au sein des données afin de détecter des anomalies.  | <input type="checkbox"/> J'ai utilisé la méthode <code>pivot_table()</code> pour créer un jeu de données avec : <ul style="list-style-type: none"> <li><input type="checkbox"/> comme lignes : des pays ;</li> <li><input type="checkbox"/> comme colonne : des indicateurs ;</li> <li><input type="checkbox"/> comme valeurs : des résultats d'agrégations d'années.</li> </ul> |  |
|   | <input type="checkbox"/> Je suis satisfait de mon interprétation des résultats de la méthode <code>describe()</code> de Pandas.<br><input type="checkbox"/> J'ai rendu visuelles et interprété les distributions d'une variable quantitative en utilisant Seaborn ou Matplotlib.   |  |
|   | <input type="checkbox"/> J'ai nommé mes variables en respect des bonnes pratiques.<br><input type="checkbox"/> J'ai implémenté au moins une fonction pour automatiser un traitement présent plusieurs fois dans le code.<br><input type="checkbox"/> J'ai utilisé au moins une boucle <code>for</code> pour éviter de répéter manuellement un traitement/une analyse.            |  |
| <b>Exercice - Partie 3</b>  |  |  |
| Configurer l'environnement de travail nécessaire à l'exploitation des données | <input type="checkbox"/> J'ai installé Poetry.<br><input type="checkbox"/> J'ai créé et rempli un fichier <code>pyproject.toml</code> en utilisant <code>poetry init</code> et/ou <code>poetry add</code> .<br><input type="checkbox"/> Le code de mon Notebook fonctionne correctement au sein de l'environnement virtuel que j'ai créé.  |  |

|                            |  |  |
|----------------------------|--|--|
| Présenter ses<br>résultats | <ul style="list-style-type: none"> <li><input type="checkbox"/> Je suis satisfait de mon rappel du contexte métier au centre de toutes les analyses.</li> <li><input type="checkbox"/> Je présente de manière précise le contenu de chaque fichier de données.</li> <li><input type="checkbox"/> Je présente mes démarches pour nettoyer le jeu de données et réduire le nombre d'indicateurs, tout en précisant les hypothèses sous-jacentes.</li> <li><input type="checkbox"/> Je suis satisfait de mon utilisation des graphiques et ai inclus des axes et des titres lisibles pour présenter mes analyses statistiques quand cela était pertinent.</li> <li><input type="checkbox"/> Je me suis efforcé de clarifier le lien entre mes résultats et l'objectif métier pour un interlocuteur de faible niveau technique.</li> <li><input type="checkbox"/> Je suis satisfait par le storytelling de ma présentation.</li> </ul> |  |
|----------------------------|--|--|