**Mise en situation pour le poste de Data Analyst**

**Introduction**

Bienvenue dans cet exercice de mise en situation pour le poste de Data Analyst chez CVE Group. Vous allez travailler avec des données provenant de **deux centrales solaires**.

* La première a une puissance crête de **10 377,6 kW** et utilise **trois pyranomètres** pour mesurer l’irradiation à différents endroits de la centrale.
* La deuxième a une puissance crête de **2 207,48 kW** et utilise **deux pyranomètres**.

Chacune des deux centrales dispose également d’un système de mesure de la production d’énergie.

Les données que vous allez analyser incluent les mesures prises toutes les 5 minutes pour chaque centrale. Elles incluent les relevés des **trois pyranomètres pour la première centrale** et des **deux pyranomètres pour la deuxième centrale**, ainsi que la production d’énergie mesurée.

**Important** : Les fonctions que vous allez écrire pour effectuer les analyses doivent être conçues de manière à être **réutilisables pour les deux centrales solaires**. Cela signifie que tout traitement ou analyse appliqué à une centrale devra pouvoir être appliqué à l'autre sans modification majeure du code.

L’exercice peut être réalisé dans le langage de programmation de votre choix.

**Jeu de données**

Le fichier CSV annexé contient les colonnes suivantes :

* **datetime** : L’horodatage des mesures.
* **pyrano1** : Mesures d’irradiation du premier pyranomètre.
* **pyrano2** : Mesures d’irradiation du deuxième pyranomètre.
* **pyrano3** : Mesures d’irradiation du troisième pyranomètre (si applicable).
* **production** : Production d’énergie mesurée sur le site.

**Objectifs de l’exercice**

1. Charger les données depuis le fichier CSV fourni.
2. Donner des informations générales sur l’échantillon de données.
3. Retirer les valeurs aberrantes des différents capteurs (production et irradiation).
4. Nommer si possible chaque défaut trouvé dans le jeu de données.
5. Calculer le pourcentage de données aberrantes par rapport au jeu de données total.
6. Ramener tout le jeu de données au pas horaire. C’est-à-dire regrouper les mesures prises toutes les 5 minutes en valeurs horaires.

**Important** : Soyez vigilant concernant le **type d’agrégation à utiliser** lors de la conversion au pas horaire. Certaines mesures peuvent être **moyennées**, tandis que d'autres, peuvent être **sommées**.

1. Calculer une matrice de corrélation entre les 3 pyranomètres et la production.
2. Identifier le pyranomètre qui corrèle le plus avec la production et expliquer ce que cela signifie.
3. Calculer le Performance Ratio (PR) au pas mensuel du site. Vous pourriez utiliser la formule du PR qui est :