**Pourquoi cette section ?**

Vous avez été nombreux à me demander du contenu sur la science des données (data science).

La science des données est un sujet extrêmement vaste et complexe qui pourrait nécessiter une formation complète de 40 heures (voire plus).

Cependant, il y a déjà beaucoup de choses que l'on peut faire avec les bases de la bibliothèque pandas, qui est la bibliothèque la plus utilisée dans ce domaine.

Lecture d'un fichier CSV ou Excel, affichage des données, traitement des valeurs, affichage des moyennes, des valeurs minimales, maximales, filtrage des données selon une condition, affichage de graphiques... Toutes ces opérations peuvent vous apporter énormément d'informations sur vos données et sont très simples à réaliser.

Dans les prochaines parties, j'ai donc essayé de vous faire un condensé de ce qui vous sera le plus utile pour le traitement de données avec pandas. J'ai appliqué la loi de Pareto. Avec 20% des fonctionnalités de pandas que l'on va voir dans cette section, vous devriez être en mesure de réaliser 80% des tâches courantes de la science des données.

La Science de Données très souvent nommer avec le terme anglais Data Science, est une rencontre de deux termes le Big Data et l’Analyse. La Data Science n’est rien d’autre qu’une science qui nous permet d’analyser les données pour en tirer des conclusions.

Quelques exemples dans la Data Science

1. Sur un site internet :

* D’où viennent les nouveaux utilisateurs ?
* Quelles sont les vidéos les plus regardées ?
* Quelle est la journée avec le plus de connexions dans la semaine ?
* Combien d’utilisateurs se sont connectés dans les 30 derniers jours ?

1. Sur Netflix

* Quel est le pays avec la plus forte croissance en nombre d’utilisateurs ?
* En moyenne, combien de temps les utilisateurs restent-ils abonnés ?
* Quels sont les genres de films les plus populaires ?
* Quel est le pays dans lequel les utilisateurs dépensent le plus ?

Procedure

1. On commence par installer jupyterlab avec pip puis on installe pandas
2. On lance jupyterlab avec la commande ***python -m jupyterlab***

Générateur de données aléatoire <https://www.mockaroo.com/>

**Installation de matplotlib**

Pour pouvoir tracer des graphiques dans un notebook jupyter, vous devez installer la bibliothèque matplotlib.

Vous pouvez pour ce faire utiliser pip, à l'intérieur de votre environnement virtuel avec pip install matplotlib (ou py -3.9 -m pip install matplotlib sur Windows) ou directement à l'intérieur d'un notebook jupyter :

1. import sys
2. !{sys.executable} -m pip install matplotlib

**Questions pour cet exercice**

Comment récupérer la durée moyenne des films présents dans le jeu de données Netflix ?

Vous devez récupérer la durée moyenne des films présents dans le jeu de données Netflix. Vous devez ignorer les rangées qui contiennent un numéro de saison en durée et ne conserver que les rangées avec des données en minutes.

Réponse

df\_duration = df[df["duration"].str.contains("min")]["duration"]

df\_duration.str.replace(" min", "").map(int).mean()