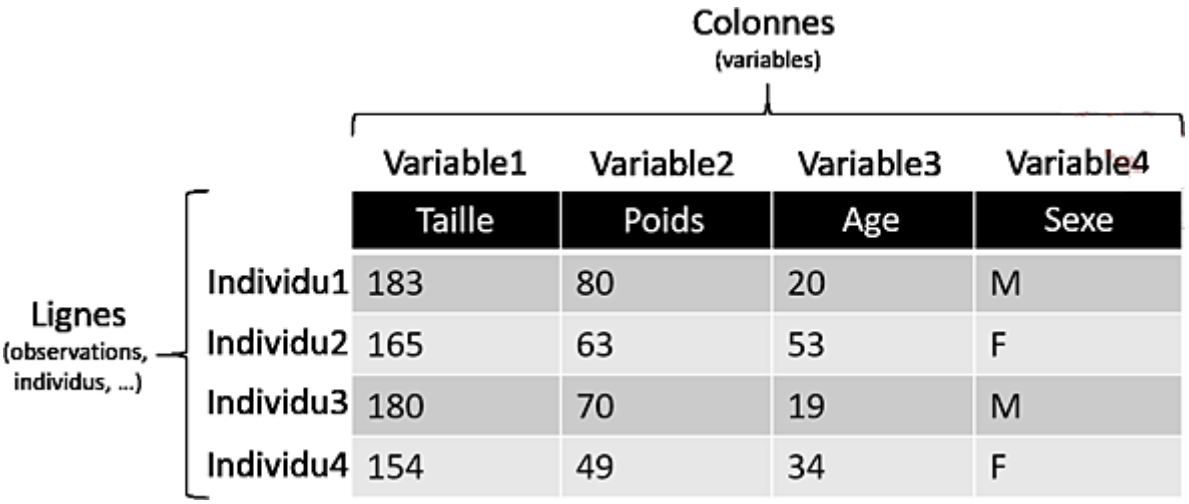
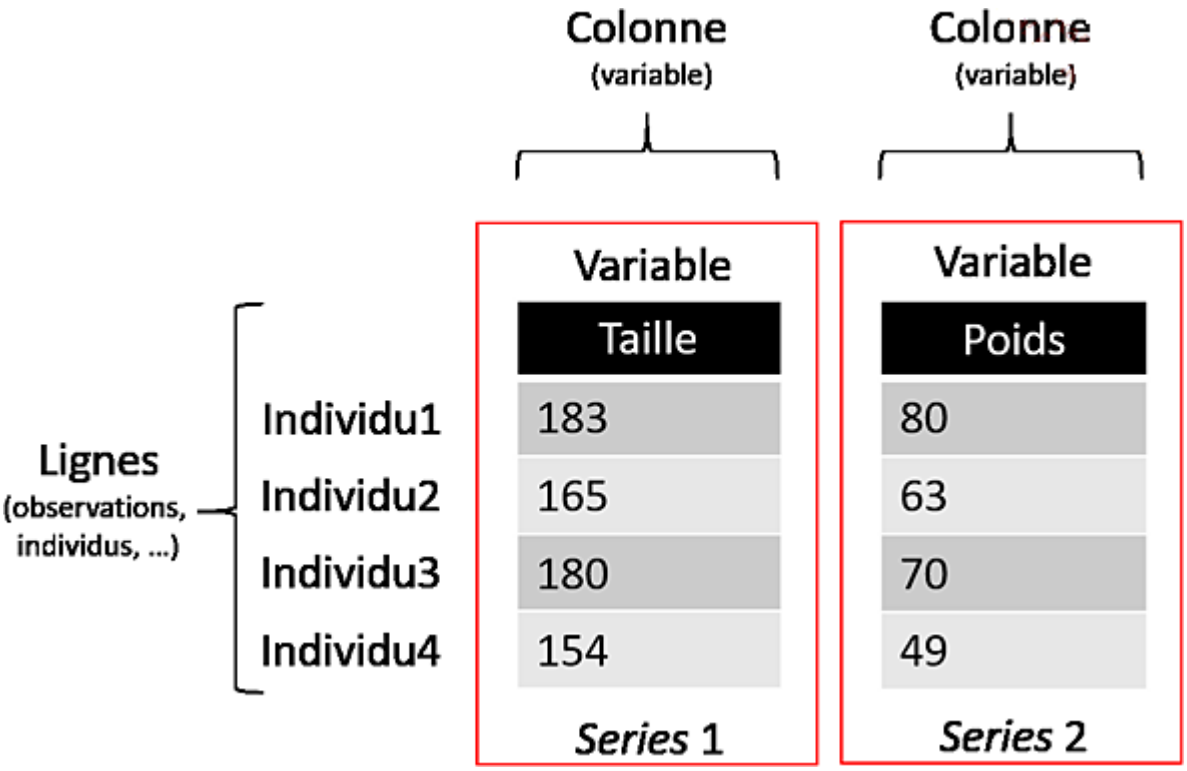
Introduction à la librairie Pandas

Pandas est une librairie Python dédiée à la Data Science (science des données, en français). Il s'agit d'ailleurs de la librairie Python la plus populaire et la plus performante pour faire de l'analyse de données.

Cette librairie amène avec elle deux nouvelles structures essentielles pour l'analyse de données, appelées **Series** et **DataFrame**. Un objet de type DataFrame, ou plus vulgairement un dataframe, peut être assimilé à un tableau à deux dimensions ou encore une feuille Excel, ce qui correspond à la structure de données la plus utilisée en Data Science. Cette structure est composée de lignes représentant des observations ou individus et de colonnes correspondant aux variables décrivant ces observations/individus.



Un objet de type "Series", ou plus simplement une série, peut être assimilé à un vecteur, c'est-à-dire à une suite de valeurs. La série correspond aussi à une colonne d'un dataframe. En réalité, le dataframe est constitué d'autant de séries qu'il a de colonnes. Ci-dessous, voici comment des séries peuvent être représentées visuellement.



C'est seulement avec l'arrivée de Pandas et ses structures de données que Python a réellement commencé à être utilisé pour faire de l'analyse de données. Cette librairie apporte non seulement de nouvelles structures ultra performantes pour stocker des données, mais elle apporte aussi un ensemble de méthodes et fonctions associées, permettant d'explorer les dataframes et séries, de les nettoyer, transformer, manipuler ou encore de les visualiser de manière très efficace et rapide.

Il faut savoir que la librairie **Pandas s'appuie fortement sur la librairie NumPy**, puisque ses structures de données sont basées sur les tableaux NumPy (ndarrays). Ainsi, en plus de la structure ndarrays, la librairie Pandas profite aussi des performances de calcul de NumPy.

L'avantage de Pandas par rapport à NumPy est que **l'objet de type DataFrame permet de stocker des données de types différents**, donc des données hétérogènes. Cela signifie que dans un dataframe, on pourra avoir une colonne contenant des chiffres, une autre contenant du texte, une autre des booléens, etc. Cette possibilité de stocker des données hétérogènes vient du fait qu'un dataframe est constitué d'un ensemble de séries qui sont indépendantes entre elles et peuvent donc contenir des types de données différents de l'une à l'autre. En revanche, au sein d'une colonne, les données doivent être de même type.

De plus, les dataframes et les séries sont fournis avec **la possibilité d'assigner des étiquettes aux données, plutôt que de travailler avec des index numériques**, comme c'était le cas avec les ndarrays de NumPy. Avec Pandas, les lignes et les colonnes peuvent être identifiées avec des étiquettes plutôt que des nombres.

Enfin, Pandas permet de manipuler efficacement les données manquantes ainsi que les données de séries temporelles (time series data), deux types de données souvent rencontrées en Data Science et dont la manipulation serait complexe sans Pandas.

Vous l'aurez compris, les séries et les dataframes sont des structures bien mieux adaptées au travail de l'analyste de données que ne le sont les ndarrays de NumPy. Ainsi, pour pouvoir faire de l'analyse de données avec Python, il est impératif de maîtriser cette librairie essentielle qu'est Pandas.