Descriptif du projet fin étude Data Science

Création d’une base de données MYSQL afin de conserver les données chargées depuis le data Set du gouvernement sur les entreprises.

Changement utilisation d’un data Set car les API ne sont pas facilement accessible.

<https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/>

Nous utiliserons le fichier StockUniteLegale\_utf8 en date du 1er Juillet 2025.

On peut trouver la documentation avec ce lien : <https://www.sirene.fr/static-resources/documentation/v_sommaire_311.htm#14>

Le fichier est très lourd (plus de 3 Go après décompression). La taille du fichier n’est pas passable sur Databricks (Impossible d’importer un fichier de plus de 500 mb).

Je décide donc de trier mes données en « chunkant » avec pandas.

Le fichier comporte l’intégralité des sociétés comprenant des entrepreneurs individuels, des PME, ETI et grandes entreprises.