TP qualité des données

Le dataset contient des données issues d'un scraping d'un site de voitures d'occasion. Le scraper va continuer de fonctionner en continu.

Il vous est demandé de réaliser une pipeline de qualité des données. L'objectif est d'automatiser le traitement des données fournies par le scraper en temps réel pour les stocker dans une base de données avec la meilleure qualité possible.

Le CSV qui vous est donné est un échantillon des données produites par le scraper. Servez-vous de ce CSV pour identifier les traitements à effectuer pour augmenter au maximum la qualité des données.

La démarche typique de qualité des données ainsi que les livrables attendus pour ce TP sont résumés ici :

1) Profiling des données avec Openrefine
    a. Identifier les données aberrantes, définir les opérations à réaliser pour les traiter. Clustering sur les colonnes pour lesquelles il est pertinent de le faire.
    b. Fournir un rapport de profilage
        i. Quelles sont les données ? Quels sont leurs intervalles ? Quelles données aberrantes trouvées ? Quels traitements à effectuer ? ...
2) Nettoyage avec pandas. On veut pouvoir :
    a. Facilement filtrer sur les marques, modèles et options, ce qui requiert de les scinder
    b. Filtrer sur le nombre de portes et le nombre de sièges
    c. Indexer sur l'année de production de chaque voiture
    d. N'avoir aucune donnée vide et aucune entrée ne doit être supprimée
    e. Mettre ces fonctions de traitement sous la forme de pipelines pandas
    f. Instancier une BDD avec le dataset nettoyé
    g. Fournir un rapport résumant cette démarche, les scripts pandas ainsi que le dataset nettoyé
3) Mise en place de tests automatisés
    a. Rédaction de tests soda-core
    b. Création d'une pipeline Dagster

Vous serez évalués sur votre démarche et sur le bon fonctionnement de la pipeline Dagster.