Travail Dirigé 2 - Travailler avec les dataframes

Nicolas Martin - Spark core batch

Objectif : Maîtriser les traitements avec les dataframes, les jointures et les udf. Utiliser à la fois le DSL et le SQL pour effectuer les traitements.

Datasets utilisés :

- Codes postaux - laposte_hesasmal.csv - https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/
- Départements français - departement.csv - https://sql.sh/1879-base-donnees-departements-francais

Questions :

1. Ecrivez une fonction (udf) qui ajoute une colonne contenant des code postaux en une colonne contenant le numéro de département.
2. Ajoutez une colonne avec le code postal utilisant cette fonction nouvellement codée.
3. Quel département contient le plus de codes postaux ?
4. Combien de code postaux un département a-t-il en moyenne ?
5. Lisez le nouveau fichier. Que pouvez vous en dire ?
6. Joignez les deux fichiers. Quel type de jointure sera exécutée ? Pourquoi ?
7. Écrivez le dataframe en fichier csv avec ces informations. Que remarquez vous ?
8. Faites la même chose en utilisant le langage SQL plutôt que le DSL. Ecrivez ici la ou les requêtes utilisées.