

TD 4 — Notes pour le chargé de TD

Analyse des données — L3 Économie — document interne

Ce TD remplace un cours. La fusion de tableaux ne sera pas présentée en cours magistral. **C'est vous qui l'introduisez**, en vingt minutes, avant que les étudiants ne se lancent. Le déroulement ci-dessous le prévoit. Les données sont réelles : les populations de référence 2023 de l'INSEE, un seul classeur Excel, chargé par son adresse. Tous les chiffres ci-dessous ont été vérifiés sur le fichier.

1. Les chiffres à connaître

Communes (feuille Communes)	34 904
Départements (feuille Départements)	100
Codes CODGEO à 5 caractères	34 788
Codes à 6 caractères (outre-mer)	116
Lignes avant et après la fusion	34 904
Communes appariées	34 900
Communes sans département	4
Commune la plus lourde : Ajaccio	45,3 % de la Corse-du-Sud

2. Déroulement

Durée	Contenu
15 min	Retour sur le TD 3. Insister sur le journal et sur le piège de MOTIF_ABS.
20 min	Introduction à la fusion , en projetant. Voir la section suivante.
55 min	Travail autonome sur les parties 1 à 4. Les questions 11 à 13 sont une réserve.

Priorité si le temps manque : les parties 1 et 3. La partie 2 (construire le code) peut se traiter au tableau ; la partie 4 se discute collectivement en dix minutes.

3. L'introduction à la fusion — vingt minutes

Projetez un exemple minuscule, que les étudiants peuvent vérifier de tête. Deux tableaux de trois lignes suffisent :

- un tableau de communes avec leur code département ;
- un tableau de départements avec leur nom, où il *manque* l'un des codes.

Faites alors, dans l'ordre :

1. `merge(..., how="inner")` : une ligne disparaît. Demandez laquelle, et si quelque chose l'a signalé. Rien.
2. `how="left"` : toutes les lignes sont là, avec une case vide.
3. `indicator=True` : la colonne `_merge` dit exactement laquelle ne s'est pas appariée.
4. Dupliquez un code dans le tableau de droite, et refaites la fusion : **le nombre de lignes augmente**, sans erreur.

Le message à laisser au tableau : les trois contrôles de toute fusion — compter les lignes avant et après, lire la colonne `_merge`, regarder ce qui ne s'est pas apparié. Une fusion ratée ne produit presque jamais d'erreur.

4. Les trois moments qui comptent

La clé, questions 3 à 5. `Code commune` est un entier : L'Abergement-Clémenciat vaut 1. Il faut `zfill(3)` pour obtenir 01001. Puis vérifier les longueurs : 116 codes font six caractères. Ne donnez pas la raison ; faites-la chercher. Ce sont les communes d'outre-mer, dont le code département a trois chiffres.

Les quatre orphelins, questions 7 et 8. Codes 975, 977, 978 : Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin. Ce sont des *collectivités* d'outre-mer, pas des départements. Le point à faire énoncer : ce n'est pas une erreur de fusion, c'est une information que seule `_merge` a révélée. Avec `how="inner"`, elles auraient disparu en silence.

Marseille introuvable, question 10. Marseille, Paris et Lyon apparaissent découpées en arrondissements municipaux, chacun pesant quelques pour cent de son département. Aucun ne peut arriver en tête. Faites-leur dire que la question 2 — de quoi une ligne est-elle la description — n'était pas une formalité.

5. Erreurs fréquentes

Erreur	Symptôme	Ce qu'elle enseigne
Oublier <code>header=7</code>	colonnes Unnamed	Ouvrir un fichier avant de le charger
Concaténer sans <code>zfill</code>	011 au lieu de 01001	Un identifiant n'est jamais un nombre
Ne pas vérifier la longueur	116 codes faux, invisibles	Une règle se vérifie
<code>how="inner"</code>	4 communes disparues sans signal	Garder tout, puis regarder
Chercher par nom	<i>Marseillette, Cognat-Lyonne</i>	On fusionne sur des codes

La question 13 corrige le code de l'outre-mer : pour un département à trois chiffres, on ne garde que les deux premiers. La dernière sortie du corrigé doit n'afficher que des longueurs de 5 — vérifiez-le à la première exécution.