# Choix base de données

1)

Création de table Département, Région et Commune pour respecter la 3NF

Pour éviter les redondances

Population intégrée à commune

Cependant pas population pour toutes les communes

2)

Jai fait le choix de prendre les communes répertorier dans

Référentiels géographiques

38 917 communes recensées contre 34 992 dans données communes (INSEE)

Cela s’explique qu’on parle dans un cas de données géographiques et de l’autre de données de population

Il manque aussi les communes des Territoires d’outre-mer (polynésie…) dans fichier INSEE et certaines communes ont été fusionnées avec une plus grande dans certains cas au cours des dernières années

Les 38 917 communes possèdent un code commune malgré tout, des ventes peuvent être associés à ces communes

Mais pas répertoriées par l’INSEE donc pas de code postale et de données de population

Choix discutable 🡪 présenter ce choix à Rami

3)

Intégration de tous les codes postaux aux 34 992 communes (INSEE) grâce à un fichier de correspondance de l’INSEE

J’ai fait le choix d’intégrer le code\_postal dans commune et pas dans bien pour éviter les redondances

On trouvera par contre le code\_commune dans l’adresse du bien

4) la table bien 🡪 qu’est ce qui garantie qu’une ligne correspond à un bien précis dans le fichier valeur foncière 🡪 suppression de tous les champs vides

Un bien c’est :

Une adresse 🡪 rue + code\_commune

Un type (appartement, maison)

Un nombre de piece

Une surface carrez

Une surface reelle bati

On supprime les doublons en prenant en compte ces criteres

24 biens supprimés 🡪 24 biens vendues plusieurs fois au cours du 1er semestre 2020

**A une meme adresse comment peut on etre sur que c’est le meme bien, si c’est un appartement, parfois appartement jumeaux dans la meme copropriété, les données indiquent que c’est le meme bien mais pas le meme en réalité**

C’est ces champs qui vont garantir l’unicité d’un bien 🡪 clé primaire

* La clé primaire va être en auto increment, pas la concaténation de tout ces critères
* Les champs qui vont garantir l’unicité d’une vente sont id\_bien + date\_vente + valeur\_fonciere, mais clé primaire en auto increment
* Comment va-t-on faire le lien lors de la création de la base de données ?

J’ai fait le choix de faire l’auto increment sur excel et de faire mes clés primaire et étrangères sur excel (avec les recherches)

Voir comment on aurait pu faire avec la base de données

* Type sequence pour id\_bien et id\_vente
* Pas de clés étrangère id\_bien dans vente à la création de la base, seulement apres l’intégration des données
* L’auto increment se fait dans pgadmin durant l’import des données
* Le lien se fait ensuite en liant les champs qui lient les 2 tables -> à creuser

5) Traitement des données

Dans quelles mesures doit-on nettoyer le jeu de données ?

Valeur\_fonciere

Doit-on supprimer ou modifier les outliers ?

Doit on supprimer ou remplir les valeurs manquantes par une moyenne ou autre ?

Que fait on des valeurs nulles ?

Surface

Doit-on supprimer ou modifier les outliers incohérents ?

Doit on modifier les surfaces incohérentes (par exemple bonne surface\_carrez mais surface\_reelle\_bati incohérente) 🡪 peut-on modifier ? pas de valeurs vraiment aberrantes