**Prise de note étude data set**

*Taille du df :* 37368 \* 27

Etude des Nan :

Exposition : + 75% vide -> probablement inexploitable !

Floor : idem

Lande\_size : 58% vide -> à étudier de + près

Ghg\_category et ghg\_value : même niveau de Nan (logique : pas de catégorie si pas de valeur !)

Idem pour energy performance

Nb\_bedrooms implique nb\_bathrooms

Size

Choix de traitement des valeurs manquantes :

* Moyenne générale :
* Par interpolation
* En mettant 0 (ou une valeur arbitraire ind du reste)

Réflexion sur les colonnes :

* On va supprimer Id car pas utile

On a les 3 types de colonnes : 20 float , 5 object et 2 int

**5 object** : ['property\_type', 'city', 'energy\_performance\_category', 'ghg\_category', 'expositon']

**2 int :** ['id\_annonce', 'postal\_code']

Certaines variables objet font référence à un classement (energy et ghg) alors que d’autre n’ont pas forcément de signification (city ou code postal)

Comme le data set est grand, il est compliqué de faire charger les figures !