**TO DO AU 23-09-2020**

**Df source :**

N\_Data = nielsen (8x17=126 ligne par EAN – 8 marchés x 17 KPI)

N\_France = Nielsen (1x17 KPI)

\*N\_prod = Nielsen 1 ligne par EAN (17 colonnes = 17 KPI 2019 France sans 26 valeurs)

O\_produit = Open food Facts (1 ligne par EAN13)

O\_ingredient= Open Food Facts (plusieurs lignes d’ingrédient par EAN13)

\*I\_label = xls 1 ligne par EAN (catégorie = col E ‘CATEGORIE PRODUIT’ onglet [VOLUME – HMSM]

**Df preprocessed :**

PROD = croisement N\_prod I\_label et O\_produit (complété de I\_ingrédient)

1. Preprocessing N\_data

=> garder EAN en string

=> extraire nom produit de ‘NIELSEN\_DESCRIPTION’+ vérfier bio sans les \*(code Alex)

=> créer colonne ‘SAME\_PRODUCT’ (code Alex)

=> ne pas virer les MDD et MDD 1er prix

1. Créer df N\_prod (à partir de Nielsen)
2. Import I\_label dans N\_prod: catégorisation italians dans data\_N (list(zip())
3. Créer PROD matcher les EAN de NIELSEN avec ceux de OPENFF (270/1100 à checker)
4. Analyser les manquants SI marque== + recette nielsen== + bio== + emballage + grammage !=

=> affecte un EAN13 recette similaire dans nouvelle colonne EAN13\_similar

1. Remplit les manquants de PROD avec la colonne SAME\_EAN13
2. Analyser tous les ingrédients dans O\_ingredient et les regrouper (tomate, huile, légumes, aromates, additif (amidon sucre E999 aromes)
3. Ajouter infos issues de O\_ingrédients dans O\_produit

**Ajouter infos issues de O\_ingredients dans O\_produit**

Features à créer

* Tomate total %
* Tomate fraiches %
* Tomate pulpe %
* Tomate concentré %
* Tomate double concentré %
* Viande (df ingredient)
* Fromage
* Huile
* Tomate autre % (?)
* Amidon
* Sucre
* Carotte
* Arome ( !!! si pas précisé naturel = arome artificiel)
* E999

Fonction ‘scoring’ à créer