CONCEVEZ UNE **APPLICATION** AU SERVICE DE LA **SANTÉ PUBLIQUE**

Sept 2022 – Projet 3 OpenClassRooms

Yves-A. Gagnard

CONTENU DE LA PRESENTATION

1. Contexte et objectifs
2. Nettoyage
3. Explorations initiales
4. Analyse : la France produit-elle de la nourriture de qualité ?

1.CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

CONTEXTE

- L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation
- Le jeu de données publiques Open Food Facts : Les champs sont séparés en quatre sections :
  - Les informations générales sur la fiche du produit : nom, date de modification, etc.
  - Un ensemble de tags : catégorie du produit, localisation, origine, etc.
  - Les ingrédients composant les produits et leurs additifs éventuels.
  - Des informations nutritionnelles : quantité en grammes d'un nutriment pour 100 grammes du produit.

OBJECTIF

- Trouver une idée d'application
- Charger et nettoyer les données
- Explorer les données (analyse univariée, multivariée, une réduction dimensionnelle, …).
- Evaluer la pertinence et la faisabilité de l'application.
  - L'application peut rester au stade de l'idée.

Source: Hypothèse de travail de l'étudiant

```
(320772, 162)
```

2.NETTOYAGE
2.1 CHARGEMENT

Source de données

- Format CVS, en accès publique sur le site [lien]
- Erreur de type sur le chargement de certaines colonnes, qui peut être ignorée (ce sont des colonnes texte de toute façon)
- Tableau de 320 772 x 162
- Pas de lignes en doublon complet
- 16 colonnes vides
- 76% de Nan

```
myDF = pd.read_csv('./Datasets/fr.openfoodfacts.org.products.csv', sep ='\t')
print("Données chargées")
```

```
D:\Users\yag\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeWarning: Columns (0,3,5,19,20,24,25,26,
27,28,35,36,37,38,39,48) have mixed types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

```
myDF.head()
```

	code	url	creator	created_t	created_datetime	last_modified_t
0	3087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893
1	4530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957
2	4559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957
3	16087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731
4	16094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653

5 rows × 162 columns

```
myColTypes[[0,3,5,19,20,24,25,26,27,28,35,36,37,38,39,48]]
```

```
code                              object
created_t                         object
last_modified_t                   object
manufacturing_places              object
manufacturing_places_tags         object
emb_codes                         object
emb_codes_tags                    object
first_packaging_code_geo          object
cities                            object
cities_tags                       object
allergens                         object
allergens_fr                      object
traces                            object
traces_tags                       object
traces_fr                         object
ingredients_from_palm_oil_tags    object
dtype: object
```

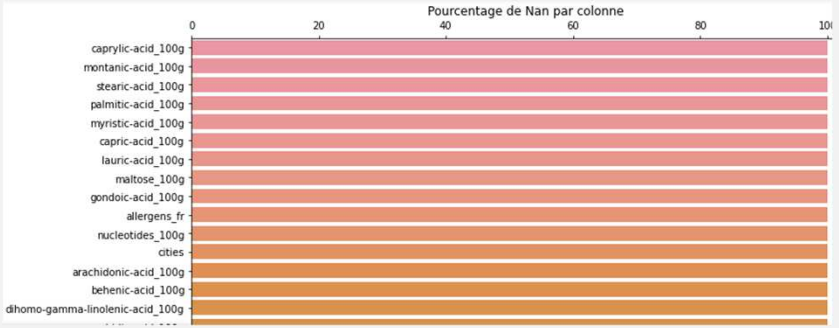
2.NETTOYAGE
2.2 DOUBLONS, NANS, VALEURS ABERRANTES

- Doublons sur une partie des champs :
  - 106095 produits apparaissent avec le même nom (« product_name ») à des centaines d'exemplaires
  - On a plus que 50 838 produits en doublon lorsqu'on exige le même nom, la même marque, le même pays.
  - C'est ceux-là qu'on retire.

```
myDF_counts_name = myDF2.product_name.value_counts()
myDF_counts_name
```

Ice Cream	410
Extra Virgin Olive Oil	303
Potato Chips	281
Premium Ice Cream	226
Tomato Ketchup	182

- Les NANs
  - Beaucoup de colonnes (~40) sont à peine remplies (>98% de Nan)
  - Ce sont sans doute car la valeur correspondante est une masse d'un ingrédient non renseignée donc valant en fait 0 (par exemple, acides rares)
  - On remplace donc ces Nan dans les ingrédients par 0.



- Valeurs aberrantes
  - On voit des valeurs aberrantes avec des ingrédients pour 100g en masse négative ou >100g
  - Après un premier essai en les ramenant à la limite (0 pour valeurs <0, 100 pour valeurs >100)…
  - Nous avons à la place utilisé un remplacement par la moyenne des valeurs les plus proches (KNN imputer)

- Valeur aberrantes (suite)
  - Apres une analyses ACP, nous avons réalisé que 1572 lignes avait un total « proteine + glucide+ lipide » > 100.
  - Nous les avons supprimés (il aurait été possible de les normaliser à 100g, mais pour si peu de ligne l'effort ne semblait pas nécessaire).

La dataframe finale est sauvée dans un nouveau fichier CSV : FoodDataNettoyee.csv

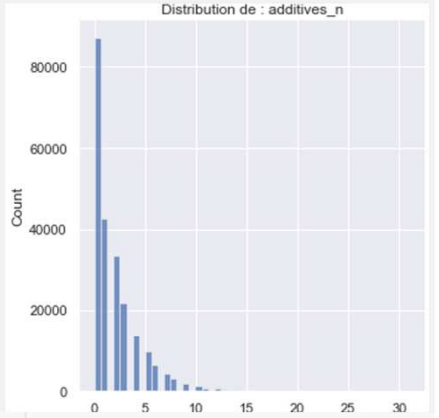
3.EXPLORATIONS INITIALES
3.1 ANALYSES UNIVARIÉES

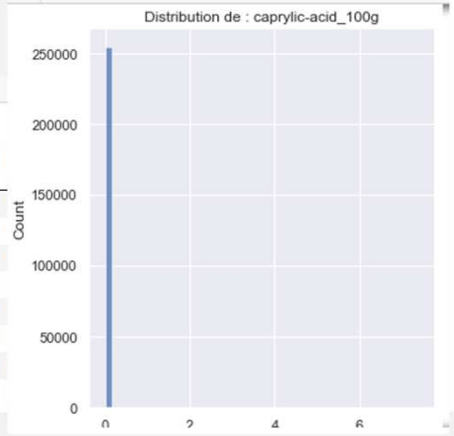
On charge les données FoodDataNettoyee.csv dans un notebook d'exploration:

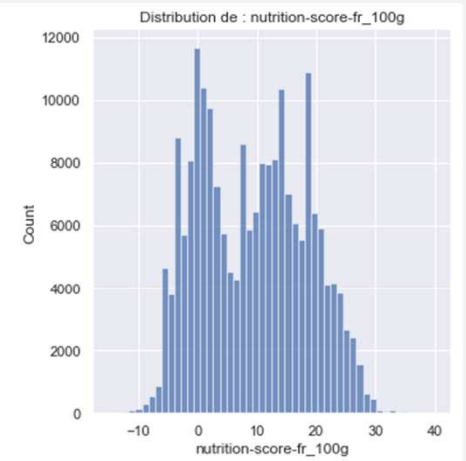
- 282 658 produit x 146 champs
- On voit 3 types de variables numériques à partir des courbes de fréquences : les décroissantes, d'autres essentiellement 0, et des variables plus ou moins centrées comme les nutriscores
- Quelques variables (ingredients_from_palm_oil_n) sont probablement des booléens (seulement valeurs 0 ou 1 ?)

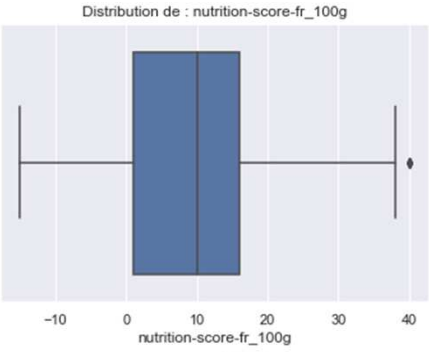
```
1 myDF[myNumericCols].describe()
```

	additives_n	ingredients_from_palm_oil_n	ingredients_that_may_be_from_palm_oil_n	energy_100g	energy-from-fat_100g	fat_100g	saturated-fat_100g
count	228995.000000	228995.000000	228995.000000	2.397150e+05	799.000000	254636.000000	254636.000000
mean	1.932623	0.019752	0.056137	1.143941e+03	568.047735	9.954727	3.813586
std	2.509797	0.140892	0.272333	6.720112e+03	693.208180	16.181286	7.204745
min	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	3.850000e+02	51.850000	0.000000	0.000000
50%	1.000000	0.000000	0.000000	1.105000e+03	300.000000	1.670000	0.180000
75%	3.000000	0.000000	0.000000	1.674000e+03	891.000000	15.040000	4.930000
max	31.000000	2.000000	6.000000	3.251373e+06	3830.000000	100.000000	100.000000



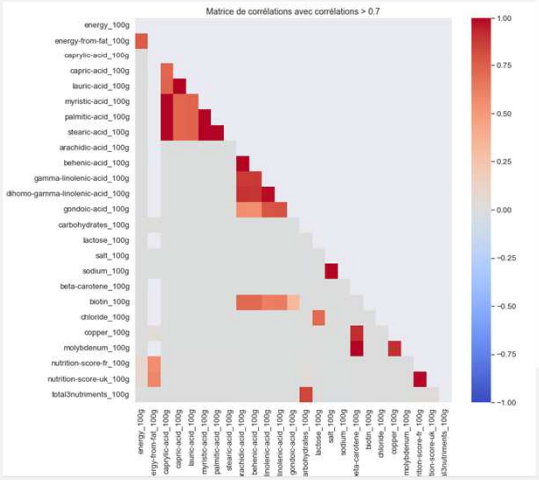


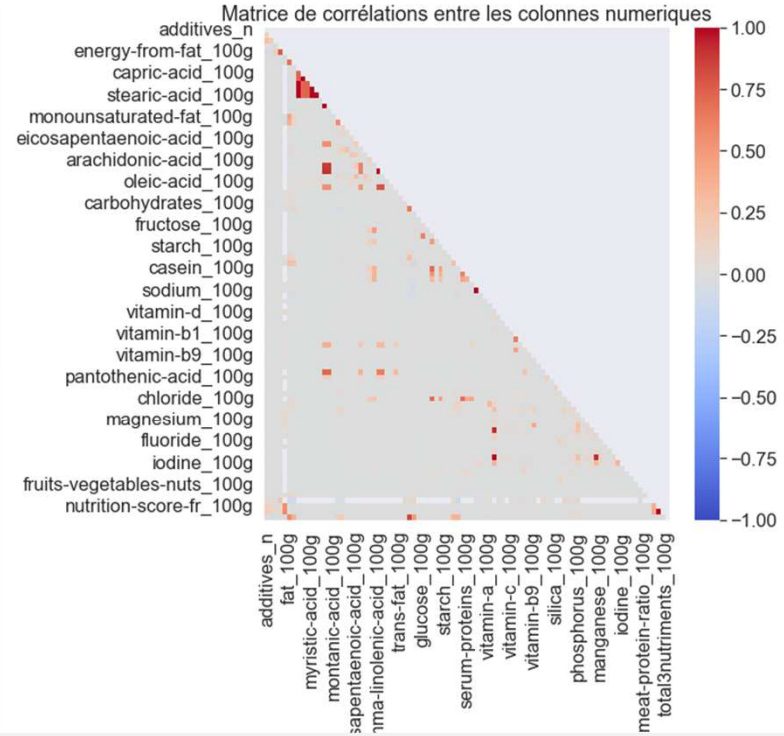




3.EXPLORATIONS INITIALES
3.2 ANALYSES BIVARIEES

- La matrice de corrélation sur l'ensemble des 91 colonnes numériques fait apparaitre certains attributs fortement corrélés (energy avec fat, certains acides, …)
- Beaucoup de corrélations sont un effet artificiel où toutes les colonnes (ingrédients) qui sont presque tout le temps égaux à zéro sont (artificiellement?) corrélés
  - la corrélation est sans doute le signe de "données détaillées non remplies" (acides et vitamines non estimés)



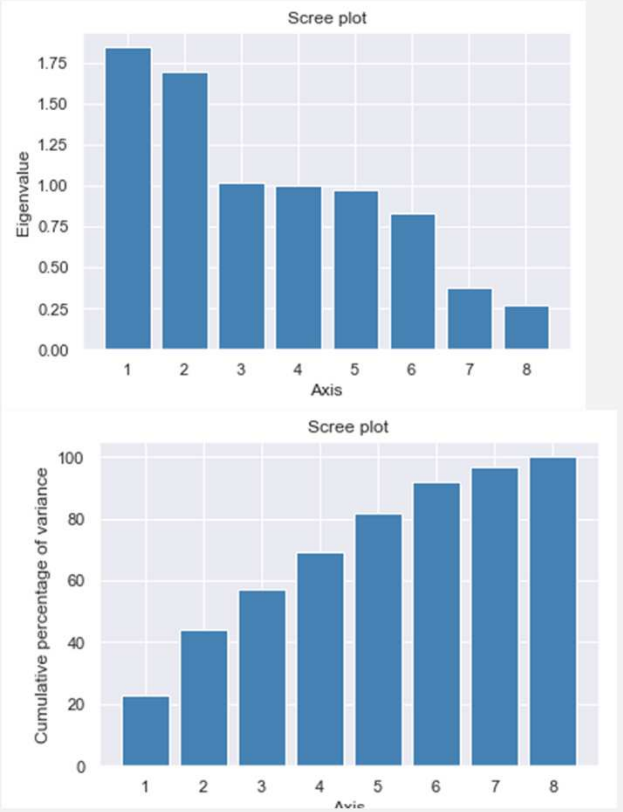


3.EXPLORATIONS INITIALES
3.3 ACP

- On a fait une première ACP à l'aide de SciKit, une seconde à l'aide de la bibliothèque fanalysis (un peu plus facile d'emploi)
- Le principe est de réduire le foisonnement des colonnes numériques (91) à 2 ou 3 dimensions.
- Pour ne pas lancer des calculs trop lourds (280k produits x 91 variables), on se limite à 8 variables (les + remplies dans le dataset initial et les + significatives):

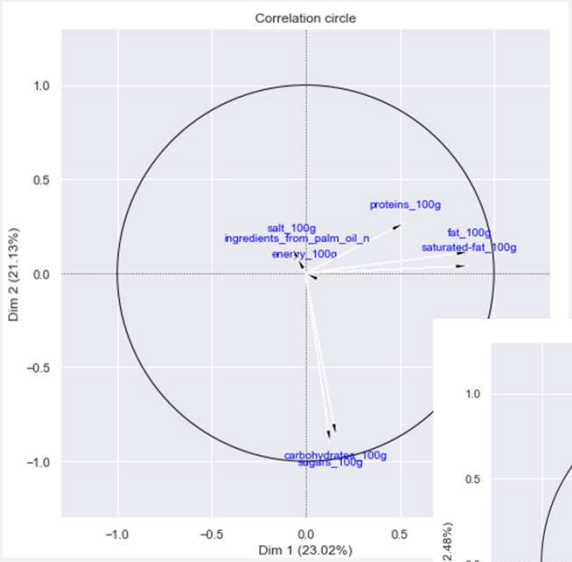
```
myColsACP = [ 'energy_100g', 'fat_100g' , 'carbohydrates_100g', 'proteins_100g', 'saturated-fat_100g',  'salt_100g',
'ingredients_from_palm_oil_n', 'sugars_100g']
```

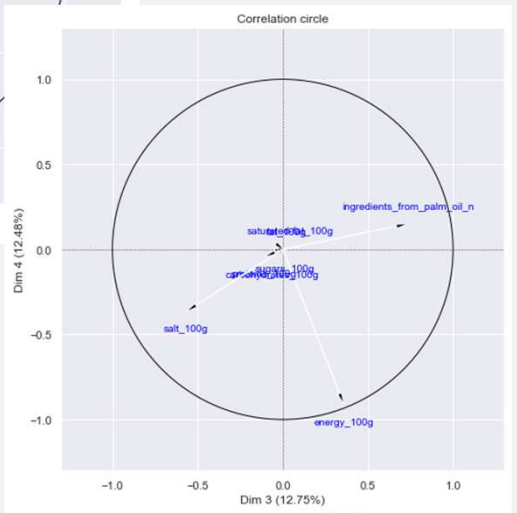
- Les éboulis montrent que le pouvoir explicatif des premières composantes n'est pas fort (<50% sur les 2 premières).
  - Explication : les 8 variables initiales sont relativement indépendantes. La réduction de dimension avec l'ACP n'est qu'un dégrossissage.



3.EXPLORATIONS INITIALES
3.3 ACP (SUITE)

- Le cercle des corrélation permet de donner une signification aux composantes :
  1. Composante 1: les produits gras et proteinés
  2. Composante 2: les produits peu sucrés
  3. Composante 3: les produits peu salés, avec huile de palme
  4. Composante 4: les produits moins énergétiques
- La composante 4 ressemble a une composante « santé », après les 3 premières qui dénotent essentiellement des ingrédients.





4. IDEE D'ANALYSE
4.1 FAUT-IL ACHETER FRANCAIS?

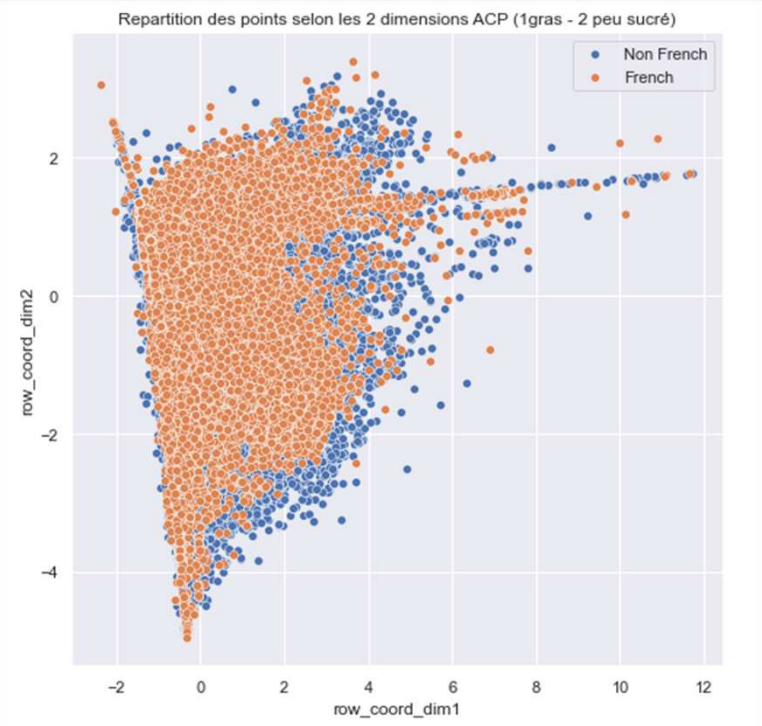
L'ECHANTILLON EST-IL SUFFISANT?
OUI

- Nb de produits nombre venant de France seule= 84 089
- venant totalement ou partiellement de France = 87 984
- Total des produits = 282 230

- La moyenne du nutriscore pour les produits d'origine française est environ 0,8 plus basse (donc meilleure) que pour les produits non français.
- Mais cela peut s'expliquer par un effet « mix » :

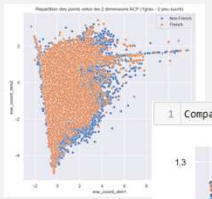
origine_france_b	nutrition-score-fr_100g
False	9.369566
True	8.598463

Y-A-T-IL UNE DIFFERENCE ENTRE LES PRODUITS FRANCAIS ET NON FRANCAIS?



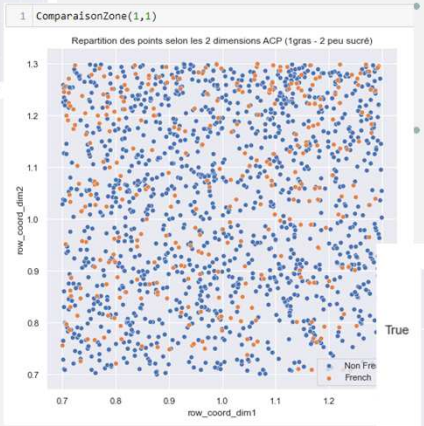
4. IDEE D'ANALYSE
4.2FAUT-IL ACHETER FRANCAIS? (SUITE)

COMPARAISON SUR CERTAINES ZONES DU GRAPHIQUE :

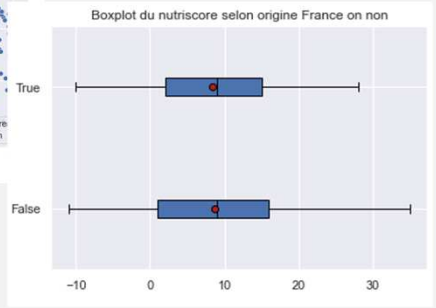


- Exemple : zoom sur la zone (1,1):

```
ComparaisonZone(1,1)
```

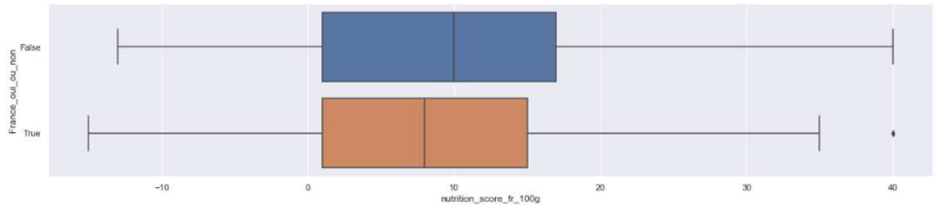


- La moyenne du nutriscore FR semble très légèrement meilleure que celle non-FR
- Il serait souhaitable de faire cette comparaison après un clustering



LA VARIANCE DES NUTRISCORE ENTRE FR ET NON-FR EST-ELLE SIGNIFICATIVE?

- Anova : p <5% donc oui, les moyennes FR et non FR ne sont pas égales.




```
import statsmodels.api as sm
from statsmodels.formula.api import ols
model = ols('nutrition_score_fr_100g ~ France_oui_ou_non', data=myDF6).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table

                                                                    érentes ou non avec :

#~H0 : Les moyennes de chaque groupe sont égales si p-value > 5%~
# H1 : Les moyennes de chaque groupe ne sont pas toutes égales < 5%
```

	sum_sq	df	F	PR(>F)
France_oui_ou_non	46,441.69	1.00	568.27	0.00
Residual	13,642,526.03	166,932.00	NaN	NaN

CONCLUSION

- Nécessaire d'arrêter les analyses à un certain point… beaucoup d'analyses supplémentaires étaient possibles :
  - Étendre l'ACP à + de variables
  - Construire un clustering (hiérarchique) à partir de l'ACP pour grouper les produits proches, ce qui rendrait plus pertinente la comparaison des nutriscores FR/non FR dans chaque cluster
  - Etude ANOVA sur les clusters les + significatifs pour valider existence d'une différence des nutriscores FR et non FR

- ⚠ Hypothèse sous-jacente qu'il aurait été souhaitable de tester au préalable : les nutriscores sont-ils bien calculés de la même manière dans tous les pays?
  - A tester avec en créant un modèle de calcul du nutriscore par régression sur les ingrédients
  - Par exemple : training sur les produits non-FR, test sur les produits FR, et on mesure si il y a une erreur systématique, ou bien régression sur les ingrédients + le pays d'origine (variable qualitative), et on regarde si le coefficient sur le pays d'origine est significatif