Santé publique France

Préparez des données pour un organisme de santé publique

Préparation et nettoyage

Imputations

Analyses uni-variées

Analyses biv-ariées

Analyses multi-variées

Intro

L’agence Santé publique France confie à votre entreprise la création d’un système de suggestion ou d’auto-complétion pour aider les usagers à remplir plus efficacement la base de données. Nous pourrions imaginer que l'utilisateur saisisse le nom ou le code barre d'un produit et que l'algorithme se charge de compléter le reste des variables. L'idée est donc de voir si un nombre limité de variables permet de déduire les autres.

Composantes **négatives** du nutriscore

Energie (KJ/100g)	Acides gras saturés (g/100g)	Sucres (g/100g)	Sodium* (mg/100g)

Composantes **positives** du nutriscore

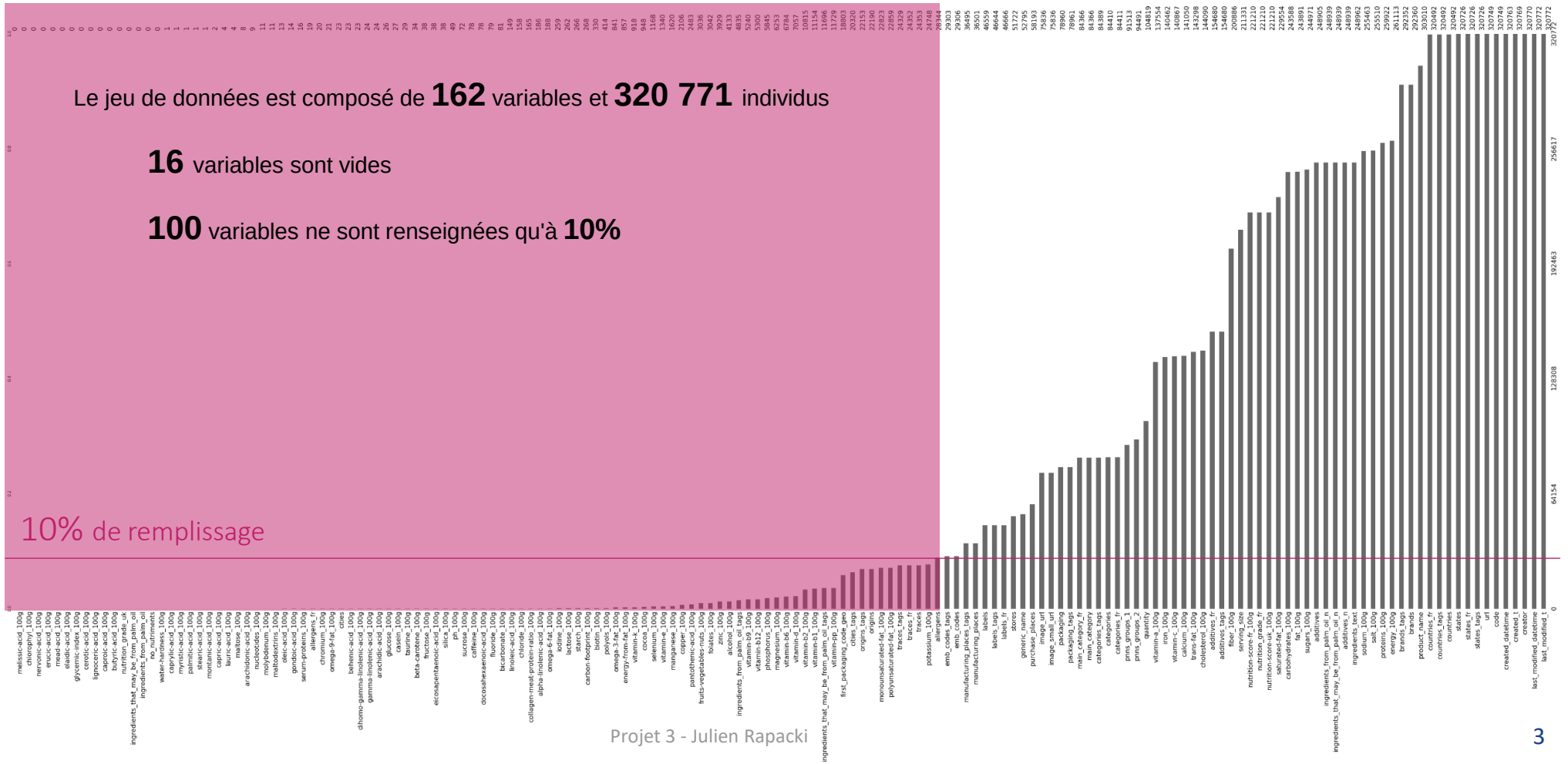
Protéines (g/100g)	Fibres (g/100g)	Fruits, légumes, légumes secs, fruits à coques, huiles de colza, de noix et d’olive* (%)

le Nutriscore varie de -15 à + 40 points

source:https://www.santepubliquefrance.fr/media/files/02-determinants-de-sante/nutrition-et-activite-physique/nutri-score/reglement-usage

Collecte des données

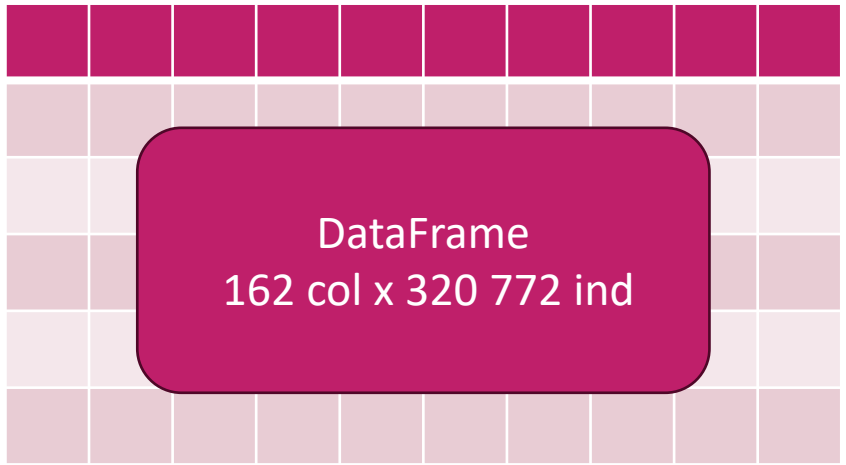
- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP



Santé publique France

Nettoyage

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

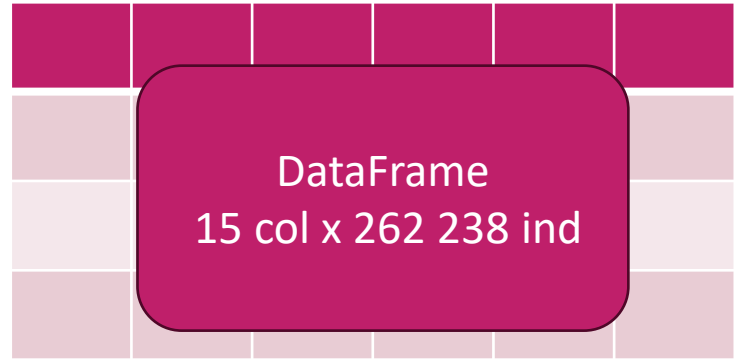


Suppressions des colonnes vides

Suppressions des colonnes inutiles dans notre contexte

Suppressions lignes en double

Suppressions des enregistrement peu renseignés*



** Lignes n'ayant aucune donnée permettant de calculer le NutriScore*

Santé publique France

Nettoyage

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

	Modalités	0	Null %
fruits-vegetables-nuts_100g	333	0.003668	0.988423
categories_fr	16437	0.000000	0.756366
main_category_fr	2374	0.000000	0.756366
pnns_groups_1	14	0.000000	0.739466
pnns_groups_2	42	0.000000	0.738692
fiber_100g	1016	0.262483	0.233955
nutrition_grade_fr	5	0.000000	0.157178
saturated-fat_100g	2197	0.262113	0.124635
sugars_100g	4068	0.141387	0.065845
sodium_100g	5291	0.130153	0.025835
product_name	187046	0.000000	0.012832
proteins_100g	2494	0.204513	0.008893
energy_100g	3997	0.033973	0.004290
countries	1030	0.000000	0.000267
code	262238	0.000000	0.000000

Les données relatives au NutriScore ont un taux de remplissage correct.
Ceci est important pour la qualité des imputations ultérieures

Santé publique France

Données brutes filtrées

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

	code	product_name	categories_fr	countries	nutrition_grade_fr	pnns_groups_1	pnns_groups_2	main_category_fr	energy_100g	saturated-fat_100g	sugars_100g	fiber_100g	proteins_100g	sodium_100g	fruits-vegetables-nuts_100g
count	262238	258873	63890	262168	221020	68322	68525	63890	2.611130e+05	229554.000000	244971.000000	200886.000000	259906.000000	255463.000000	3036.000000
unique	262238	187046	16437	1030	5	10	36	2374	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	0000000004530	Ice Cream	Snacks sucrés,Biscuits et gâteaux,Biscuits	US	d	unknown	unknown	Boissons	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	410	708	169836	62763	12872	12872	2440	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.141915e+03	5.129932	16.003484	2.862111	7.076366	0.798815	31.458587
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6.447154e+03	8.014238	22.327284	12.867578	8.409137	50.504428	31.967918
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.000000e+00	0.000000	-17.860000	-6.700000	-800.000000	0.000000	0.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.770000e+02	0.000000	1.300000	0.000000	0.700000	0.025000	0.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.100000e+03	1.790000	5.710000	1.500000	4.760000	0.229000	23.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.674000e+03	7.140000	24.000000	3.600000	10.000000	0.541000	51.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.251373e+06	550.000000	3520.000000	5380.000000	430.000000	25320.000000	100.000000

Santé publique France

Recherche de doublons

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

22 lignes en double (vraisemblablement dues à un décalage de lignes) que nous choisissons de supprimer

```
1 dataP3.code.duplicated().sum()
```

```
22
```

```
1 dataP3[dataP3.duplicated(['code'],keep=False)]
```

	code	product_name	categories_fr	countries	nutrition_grade_fr	pnns_groups_1	pnns_groups_2	main_category_fr	ener
189068	NaN	Belgique,France	6	en:to-be-completed,en:nutrition-facts-completed,en:ingredients-completed,en:expiration-date-to-be-completed,en:characteristics-completed,en:categories-completed,en:brands-completed,en:packaging-completed,en:quantity-completed,en:product-name-completed,en:photos-to-be-validated,en:photos-uploaded	NaN	NaN	NaN	NaN	
				en:to-be-checked,en:complete,en:nutrition-					

Santé publique France

Traitement des valeurs aberrantes

Variables catégorielles

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

On comptabilise les modalités

```
1 dataP3.nunique()
```

code	262238
product_name	187046
categories_fr	16437
countries	1030
nutrition_grade_fr	5
pnns_groups_1	14
pnns_groups_2	42
main_category_fr	2374
energy_100g	3997
saturated-fat_100g	2197
sugars_100g	4068
fiber_100g	1016
proteins_100g	2494
sodium_100g	5291
fruits-vegetables-nuts_100g	333

Rationalisation

```
1 print(dataP3['pnns_groups_1'].unique())
```

```
[nan 'unknown' 'Fruits and vegetables' 'Sugary snacks' 'Composite foods'
 'Fish Meat Eggs' 'Beverages' 'Fat and sauces' 'Cereals and potatoes'
 'Milk and dairy products' 'Salty snacks' 'fruits-and-vegetables'
 'sugary-snacks' 'cereals-and-potatoes' 'salty-snacks']
```

```
1 dataP3['pnns_groups_2'].unique()
```

```
array([nan, 'unknown', 'Vegetables', 'Biscuits and cakes',
       'Pizza pies and quiche', 'Meat', 'Sweets', 'Sweetened beverages',
       'Dressings and sauces', 'One-dish meals', 'Soups', 'Cereals',
       'Fruits', 'Milk and yogurt', 'Fats', 'Non-sugared beverages',
       'Cheese', 'Chocolate products', 'Sandwich', 'Bread', 'Nuts',
       'Legumes', 'Breakfast cereals', 'Appetizers',
       'Artificially sweetened beverages', 'Fruit juices', 'Eggs',
       'Fish and seafood', 'Dried fruits', 'Ice cream', 'Processed meat',
       'Potatoes', 'vegetables', 'pastries', 'Dairy desserts',
       'Alcoholic beverages', 'Fruit nectars', 'fruits',
       'Salty and fatty products', 'Tripe dishes', 'cereals', 'legumes',
```

Traitement des valeurs aberrantes

Variables numériques

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

	energy_100g	saturated-fat_100g	sugars_100g	fiber_100g	proteins_100g	sodium_100g	fruits-vegetables-nuts_100g
count	2.611130e+05	229554.000000	244971.000000	200886.000000	259906.000000	255463.000000	3036.000000
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	1.141915e+03	5.129932	16.003484	2.862111	7.076366	0.798815	31.458587
std	6.447154e+03	8.014238	22.327284	12.867578	8.409137	50.504428	31.967918
min	0.000000e+00	0.000000	-17.86	-6.70	-800. 00	0.000000	0.000000
25%	3.770000e+02	0.000000	1.300000	0.000000	0.700000	0.025000	0.000000
50%	1.100000e+03	1.790000	5.710000	1.500000	4.760000	0.229000	23.000000
75%	1.674000e+03	7.140000	24.000000	3.600000	10.000000	0.541000	51.000000
max	3.251373e+06	550.000	3520.00	5380.00	430.000	25320.000000	100.000000

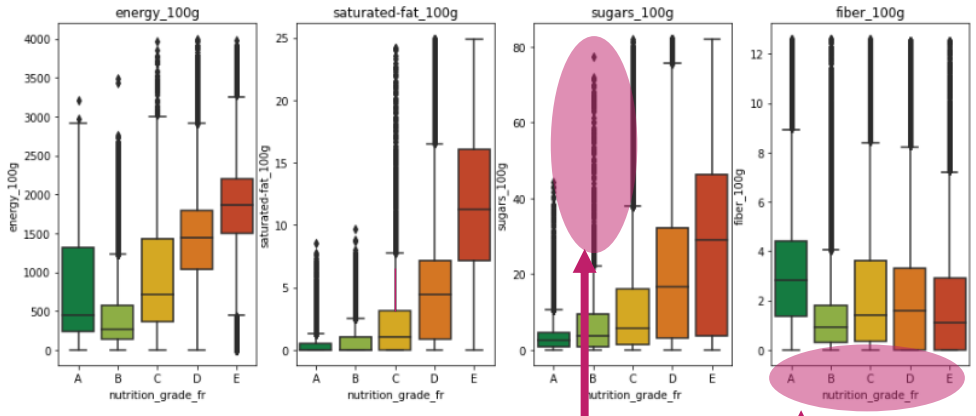
Valeurs à plus de 100g par 100g de produit
Energie > 4000 KJ/100g de produit
convertis en Nan

source https://www.fao.org/3/y5022e/y5022e04.htm

Santé publique France

Traitement des outliers

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP



	Nb Outliers sup	Nb Outliers in
energy_100g	0	0
saturated-fat_100g	4423	0
sugars_100g	3557	0
fiber_100g	5819	0
proteins_100g	2464	0
sodium_100g	10139	0
fruits-vegetables-nuts_100g	0	0

Santé publique France

Imputations - gestion des valeurs manquantes

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

```
1  dataP3.isna().mean().sort_values()
```

code	0.000000
countries	0.000267
energy_100g	0.004290
proteins_100g	0.008893
product_name	0.012832
sodium_100g	0.025835
sugars_100g	0.065845
saturated-fat_100g	0.124635
nutrition_grade_fr	0.157178
fiber_100g	0.233955
pnns_groups_2	0.738692
pnns_groups_1	0.739466
categories_fr	0.756366
main_category_fr	0.756366
fruits-vegetables-nuts_100g	0.988423

Nous nous focalisons sur les variables cibles ci-contre : Numériques et catégorielles

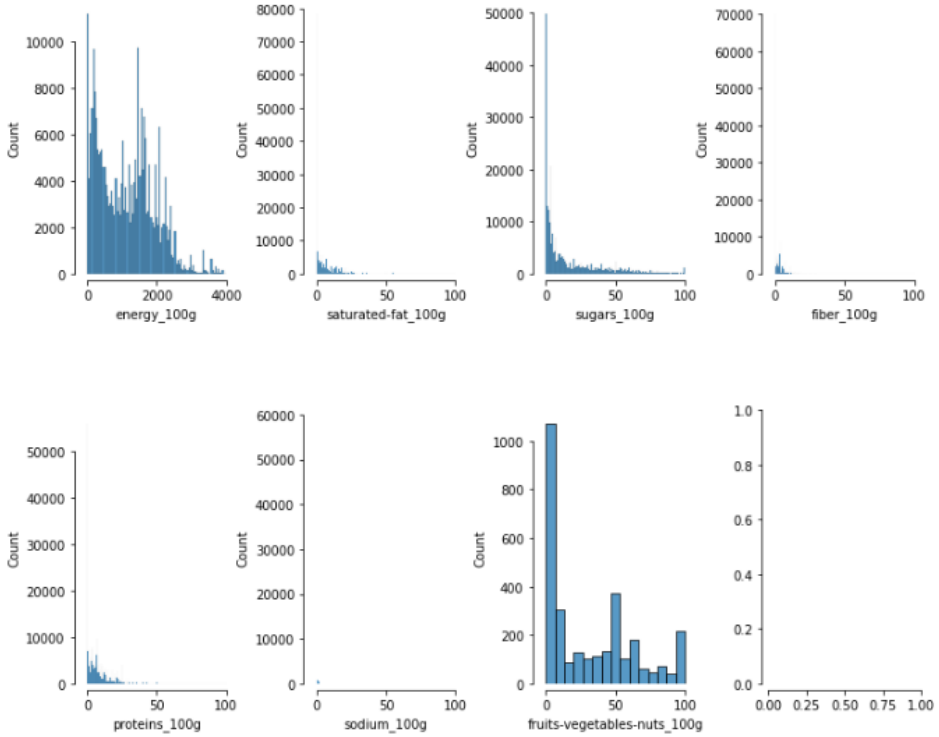
Santé publique France

Imputations - gestion des valeurs manquantes

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

Les distributions des variables numériques sont dans l'ensemble biaisées.
Les variables « energy » et « fruits-vegetables… » ont une distribution binomiale

Des imputations trop « simples » comme la médiane / moyenne sont susceptibles de manquer de précision



Imputations - gestion des valeurs manquantes

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

1 — Imputation par 0

La variable « fruits-vege… » n'est renseignée que dans 1-2% des cas

2 — Iterative Imputer

Prend en entrée les variables permettant de calculer le Nutriscore

3 — KNN

prend en entrée les variables complétées en 1 et 2 afin de déterminer les NutriScores manquants

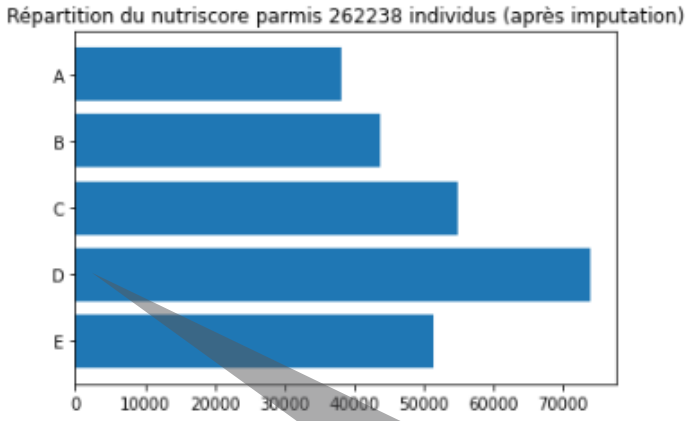
```
1 dataP3.isna().mean().sort_values()
```

	code	0.000000
	countries	0.000267
(2)	energy_100g	0.004290
(2)	proteins_100g	0.008893
(2)	product_name	0.012832
(2)	sodium_100g	0.025835
(2)	sugars_100g	0.065845
(2)	saturated-fat_100g	0.124635
(3)	nutrition_grade_fr	0.157178
(2)	fiber_100g	0.233955
	pnns_groups_2	0.738692
	pnns_groups_1	0.739466
	categories_fr	0.756366
	main_category_fr	0.756366
(1)	fruits-vegetables-nuts_100g	0.988423

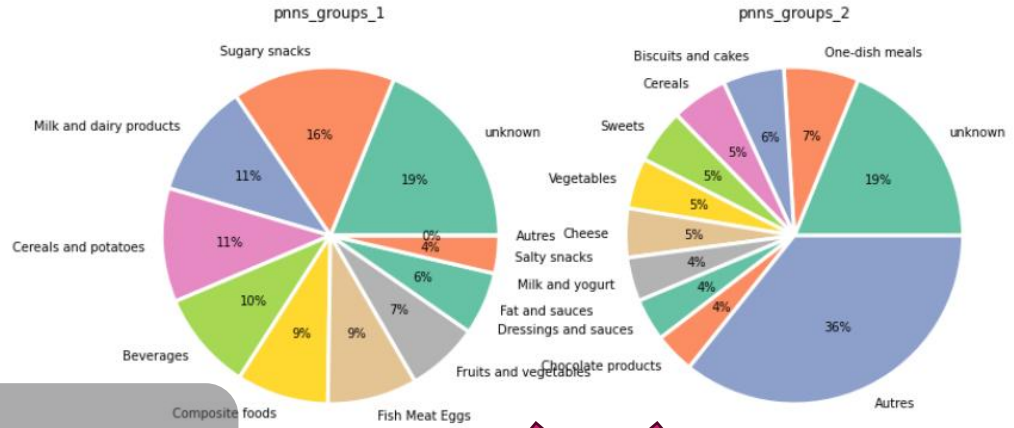
Santé publique France

Analyses univariées

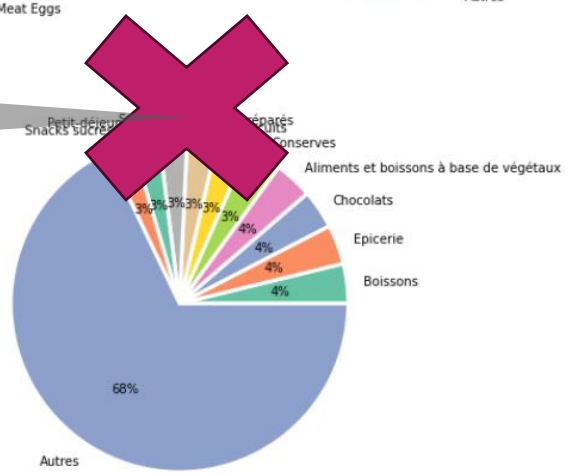
- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP



Répartition relativement homogène avec une forte représentation de la catégorie D

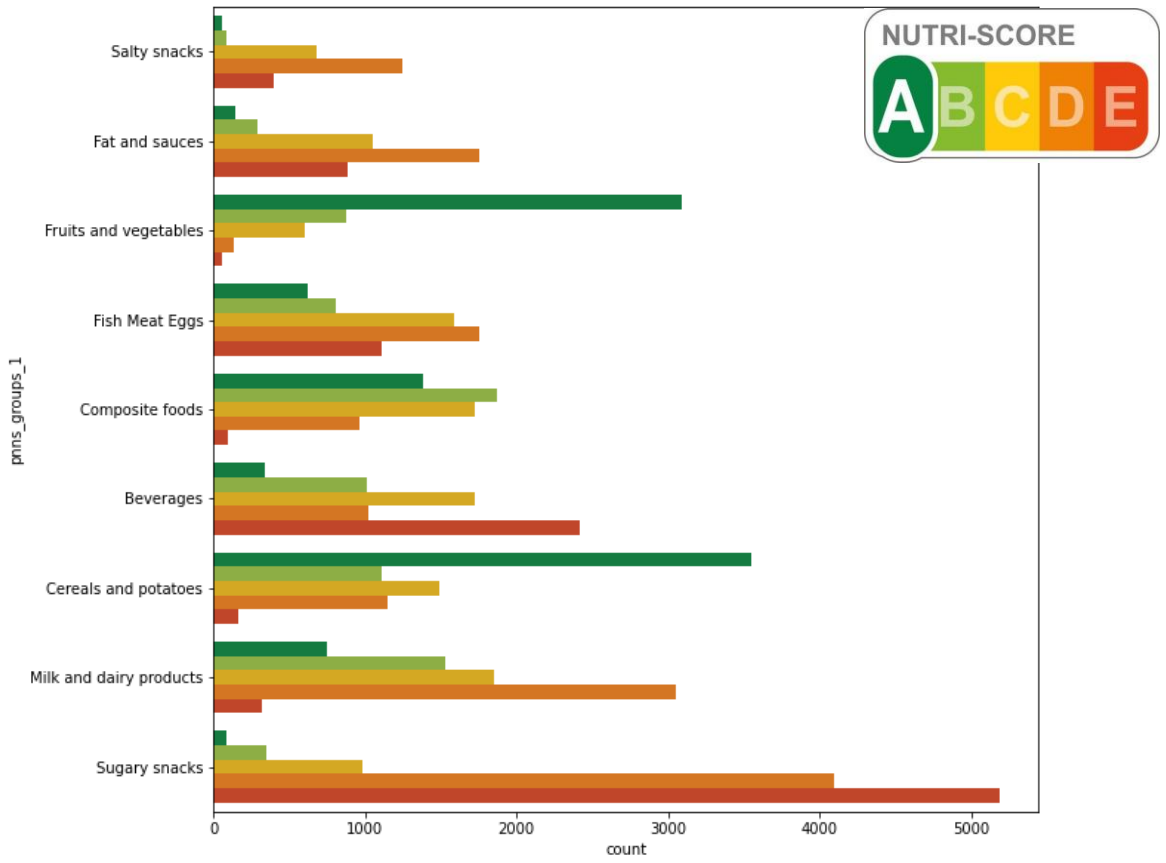


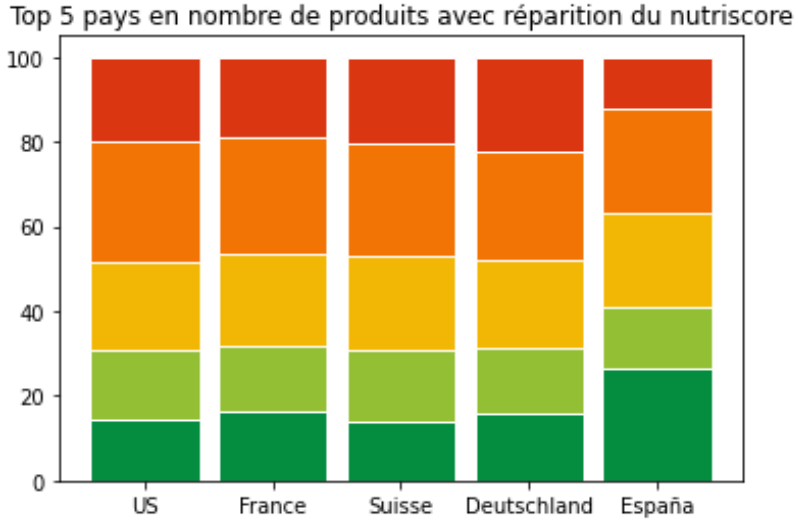
Supprimé car trop de modalités



Analyses bi-variées

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

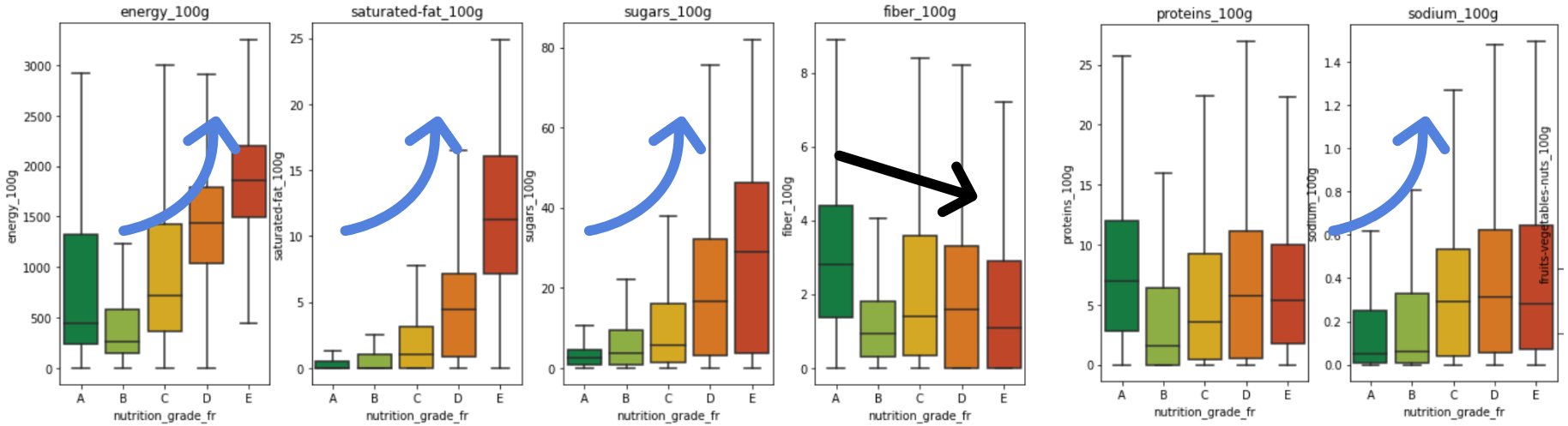




Santé publique France

Analyses bi-variées

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP



On retrouve globalement la logique de contribution positive et négative des nutriments sur les différents scores, avec de fortes amplitudes pour certains nutriments

Santé publique France

ANOVA

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

Y a-t-il une différence significative entre les différents Nutriscores A,B,C,D,E?

Hypothèse **H0** -> la différence entre les moyennes des différents groupes n'est pas significative.

Var	p_value pour la normalité
energy_100g	0.0
saturated-fat_100g	0.0
sugars_100g	0.0
fiber_100g	0.0
proteins_100g	0.0
sodium_100g	0.0

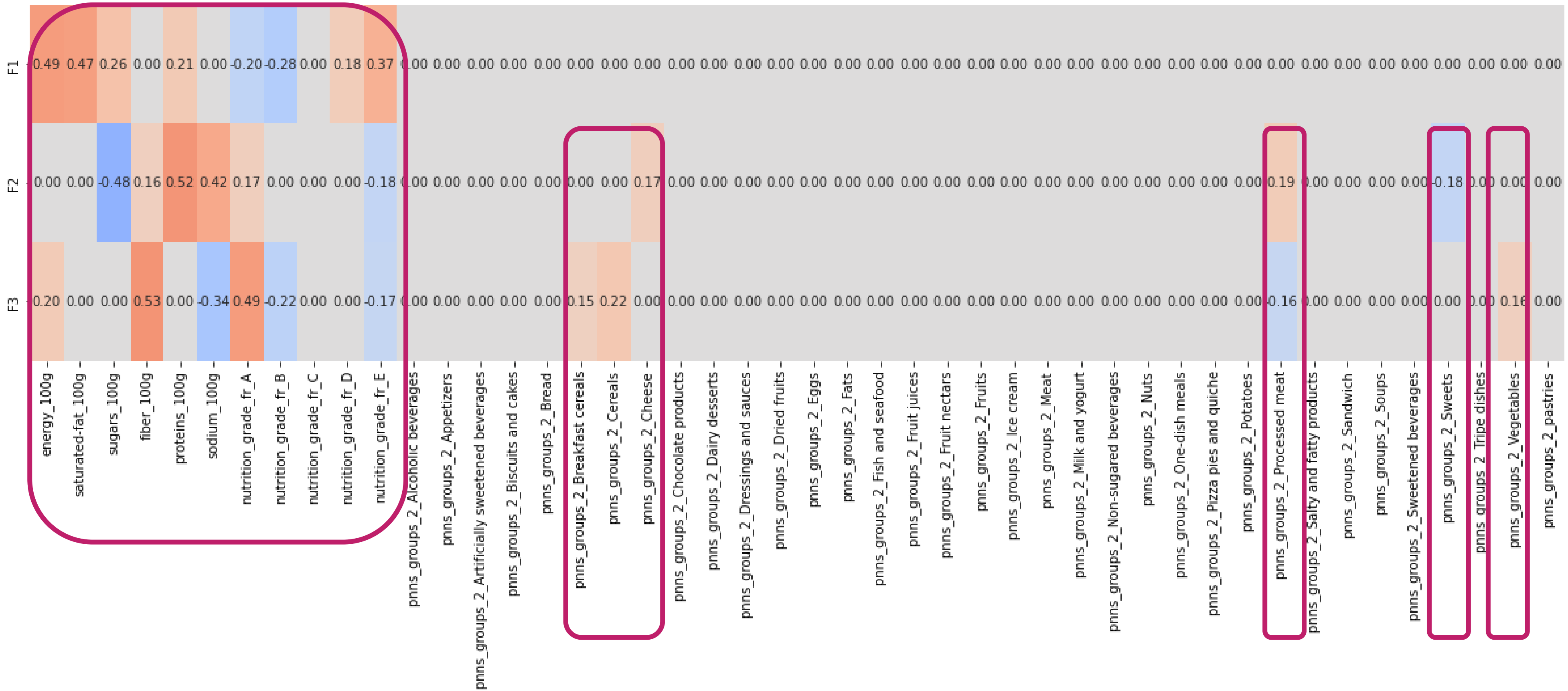
Nous pouvons rejeter l'hypothèse H0

Santé publique France

Analyse en Composante Principale

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

Il est possible de « simplifier » le jeu de données



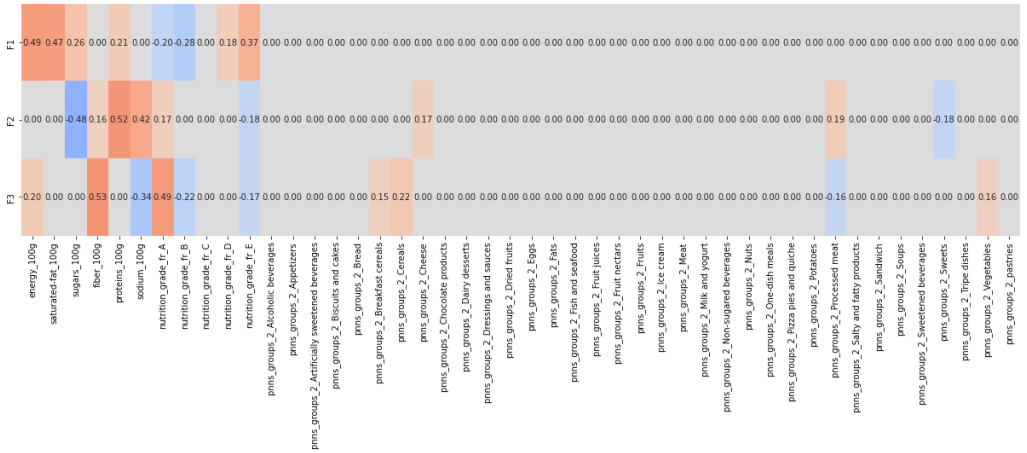
Santé publique France

Analyse en Composante Principale

- Nettoyage
- Imputations
- Analyses univariées
- Analyses bivariées
- ANOVA
- ACP

Il est possible de « simplifier » le jeu de données

F1 : « Mauvais / bon élèves »	F2 « Salé / sucré »	F3 : « Transformation»
• Fortement énergétiques • Riches en graisse et sucre • Protéines • Appartenant majoritairement au groupes D,E	• Faiblement sucrés • Riches en protéines et sel • Plutôt classés A • Produits d'origine animal « viande/fromages »	• Riches en fibres • Faibles en sel • Relativement énergétiques • Souvent classé A



Santé publique France

Rappel concernant la **Réglementation Générale sur la Protection des données.**

1 - NE COLLECTER QUE LES DONNÉES NÉCESSAIRES POUR ATTEINDRE VOTRE OBJECTIF

- ✓ Seules les données strictement nécessaires à l'élaboration des indicateurs ont été utilisées via une sélection et un nettoyage des données

2 - TRANSPARENCE

- ✓ Les données analysées ne comportent pas d'informations susceptibles s'identifier directement ou indirectement des personnes

3 - ORGANISEZ ET FACILITEZ L'EXERCICE DES DROITS DES PERSONNES

- ✓ Les sources ainsi que les coordonnées de l'auteur de l'analyse sont communiquées dans le projet

4 - DURÉES DE CONSERVATION LIMITÉES

- ✓ Ces données seront conservées uniquement le temps nécessaire à l'élaboration du projet

5 - SÉCURITÉ

- ✓ Les données sources sont du domaine publique mais les analyses qui en sont faites sont stockées sur mon ordinateur personnel portégé par un mot de passe

Santé publique France

Conclusion

Le jeu de donnée confié est perfectible et peu rendre compliqué l'élaboration d'un algorithme d' auto-complétion.

Points positifs	Points négatifs
Nombre conséquent et représentatif des produits	un certain nombre de contraintes de saisie sont nécessaires afin d'éviter les saisies aberrantes
Il y a une différence significative entre les différents NutriScore (cf ANOVA)	Données très parcellaires. Là aussi il serait pertinent de limiter le nombre de varia
la base rassemble les variables importantes	Trop de variables, ce qui engendre un fort taux de champs incomplets