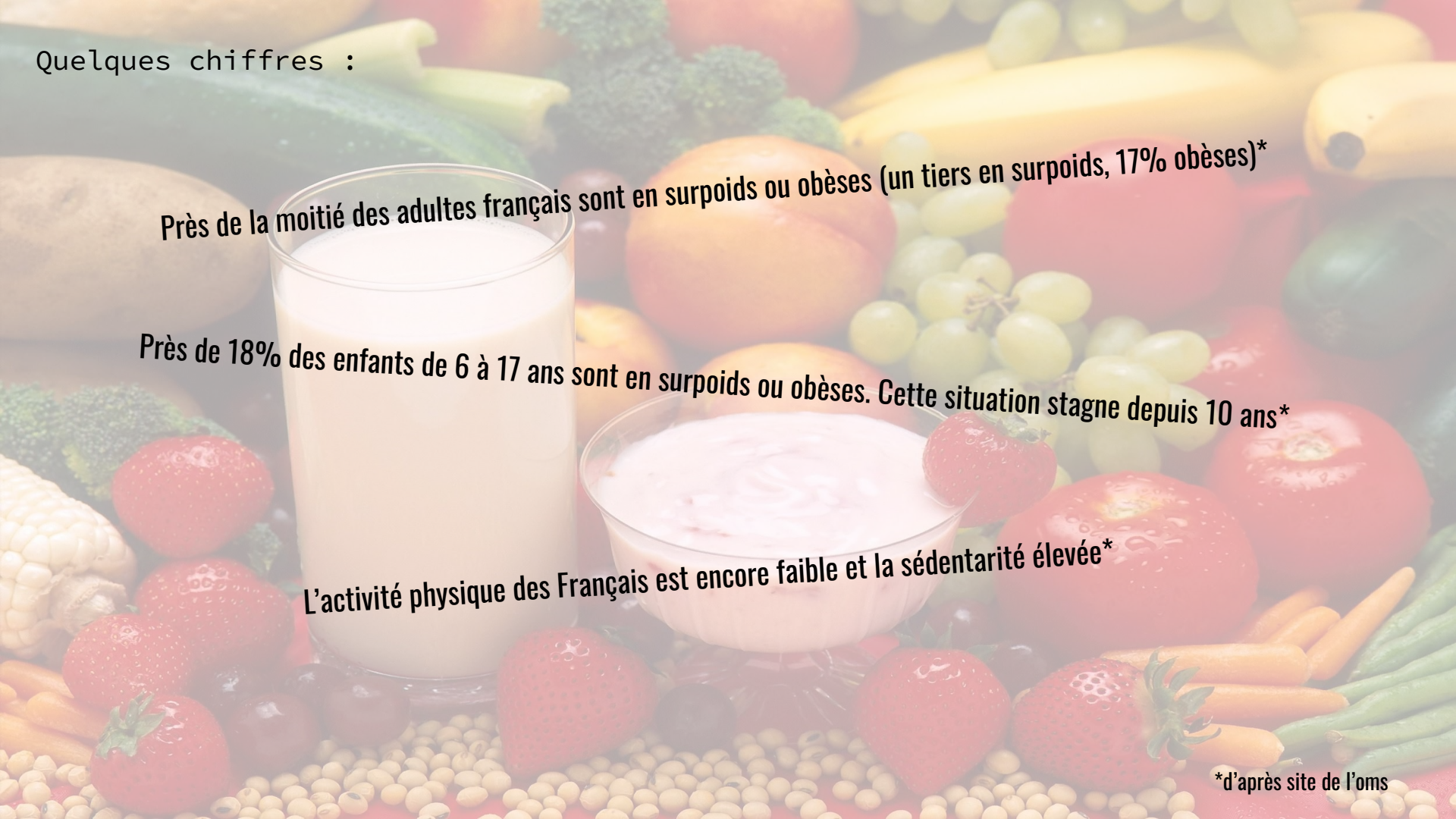


Open Food Fact

Concevez une application aux services santé public





Quelques chiffres :

Près de la moitié des adultes français sont en surpoids ou obèses (un tiers en surpoids, 17% obèses)*

Près de 18% des enfants de 6 à 17 ans sont en surpoids ou obèses. Cette situation stagne depuis 10 ans*

L'activité physique des Français est encore faible et la sédentarité élevée*

*d'après site de l'oms

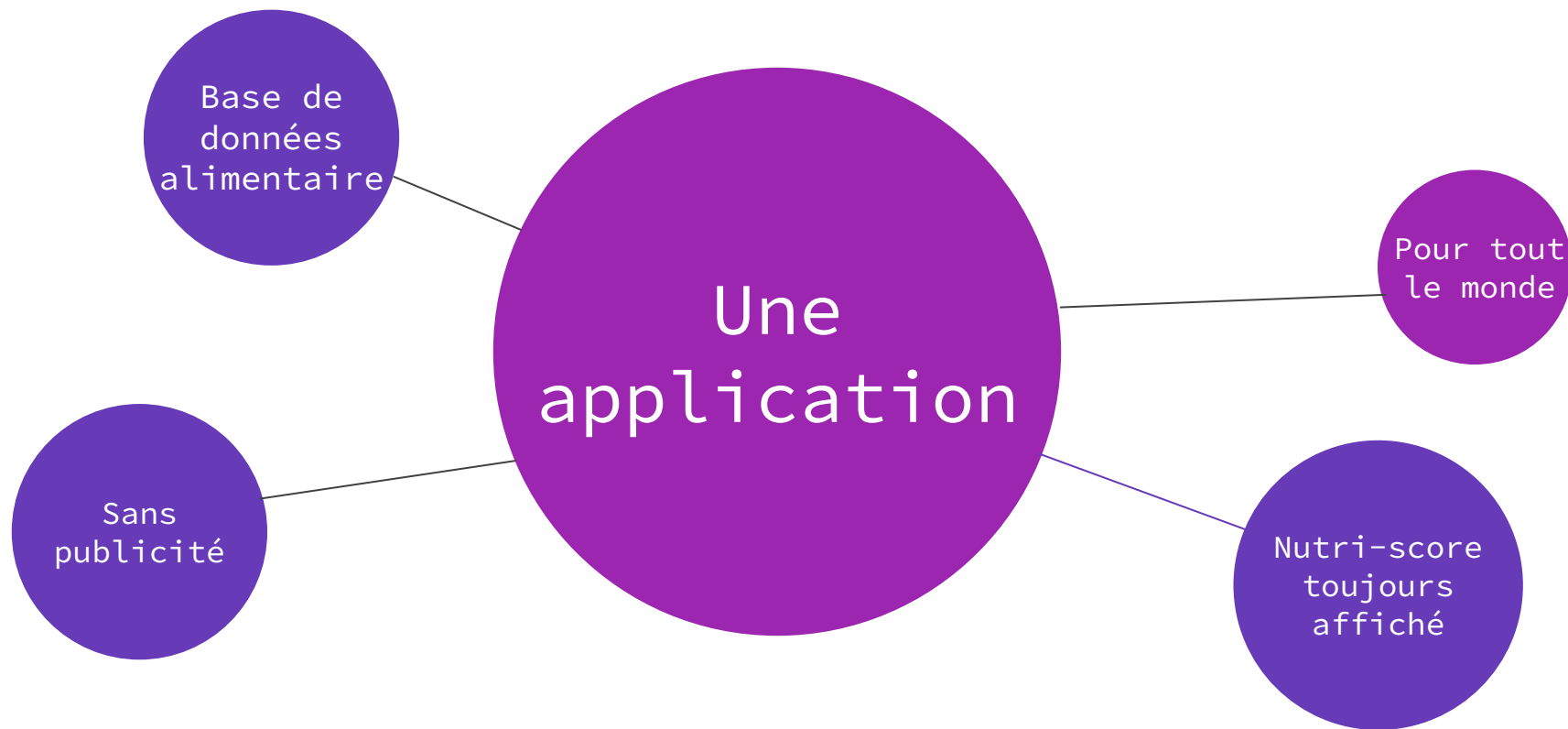
À propos



- Selon les directives gouvernementales, il ne faut pas manger trop gras, trop salé, trop sucré. Ces données seront exprimées en pourcentage de la composition et en pourcentage des AJR(Apports Journaliers Recommandés)
- On ajoutera le score énergétique. Notre référence sera les AR = Apport de Référence pour un adulte (8400KJ) ou 3700kj pour 100g
https://fr.wikipedia.org/wiki/Apports_journaliers_recommand%C3%A9s
- Ainsi que le nutri-score pour une visualisation rapide

En scannant le produit, l'application sera à la portée de tous, claire et simple d'utilisation

Ouvrez votre nourriture pour savoir ce que vous mangez



Visualisation des données

Regards sur notre jeu de données

Dans nos données, il y a (320772, 162) ce qui correspond aux lignes, et aux colonnes

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity
0	3087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN	1kg
1	4530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN	NaN
2	4559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN	NaN
3	16087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	NaN	NaN
4	16094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN	NaN

doublons sur notre jeu de données ?

code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	gener
58001	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055734	2017-03-09T10:35:34Z	1489055734	2017-03-09T10:35:34Z	Organic Salted Pistachios	
58001	http://world-fr.openfoodfacts.org/produit/0005...	kiliweb	1487432837	2017-02-18T15:47:17Z	1487432838	2017-02-18T15:47:18Z	Bramley Apple Crumble	

Des doublons ont été constaté sur notre jeu de données par rapport aux codes.



133 rows × 162 columns

Je les supprime pour avoir un code article unique.



Valeurs manquantes sur notre jeu de données

Pour le jeu de données Food, nous avons : 76.22 % de valeurs manquantes



last_modified_t	0.000000
last_modified_datetime	0.000000
creator	0.000623
created_t	0.000935
created_datetime	0.002806
code	0.007170
url	0.007170
states	0.014340
states_tags	0.014340
states_fr	0.014340
countries_fr	0.087289
countries	0.087289
countries_tags	0.087289
product_name	5.537266
brands	8.857382
brands_tags	8.859876
energy_100g	18.598568
proteins_100g	18.969860
salt_100g	20.345292
sodium_100g	20.359944
ingredients_text	22.386617
ingredients_that_may_be_from_palm_oil_n	22.393787
additives_n	22.393787
ingredients_from_palm_oil_n	22.393787
additives	22.404387
sugars_100g	23.630803
fat_100g	23.967491
carbohydrates_100g	24.061951
saturated-fat_100g	28.437021
nutrition-score-uk_100g	31.038245
nutrition-score-fr_100g	31.038245
nutrition_grade_fr	31.038245
serving_size	34.118003
fiber_100g	37.374210

Je décide de filtrer les colonnes où il y a au moins 50% de données.

Après une analyse de nos colonnes restantes, nous confirmons que notre choix d'application est approprié sur ce dataset. C'est pourquoi je sélectionne que ce dont j'ai besoin pour l'application. Je remarque la colonne 'pnns_group_1' correspondant au catégorie de produits, qui peut être intéressante.

Les catégories sur notre jeu de données

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

Après un regards sur la colonne 'pnns_groups_1', je remarque :

- Des valeurs nan: ce qui veut dire que certains produits ne sont pas catégorisé, je décide donc de les inclure dans la catégorie 'unknown'
- Je remarque aussi 2 catégories identiques mais avec une police différentes. La solution -> Je les fusionne !

sélection colonnes et lignes

```
1 def colonnes(df):  
2     features = df[['code',  
3                   'countries_fr',  
4                   'product_name',  
5                   'pnns_groups_1',  
6                   'energy_100g',  
7                   'salt_100g',  
8                   'sugars_100g',  
9                   'fat_100g',  
10                  'saturated-fat_100g',  
11                  'nutrition-score-fr_100g',  
12                  'nutrition_grade_fr']]
```



Avec ce code, je garde les colonnes indiqué en rouge.



Avec celui-ci, je garde les produits uniquement vendu en France

```
country = df[df['countries_fr'].str.contains('FR')]
```

détection de valeurs aberrantes

```
1 def valeurs_aberrantes(df):  
2     valeurs = df[(df.energy_100g > 3700) |  
3                 (df.salt_100g < 0) |  
4                 (df.salt_100g > 100) |  
5                 (df.sugars_100g < 0) |  
6                 (df.sugars_100g > 100) |  
7                 (df.fat_100g < 0) |  
8                 (df.fat_100g > 100) |  
9                 (df['saturated-fat_100g'] < 0) |  
10                (df['saturated-fat_100g'] > 100)]  
11
```

Les valeurs renseignées sont pour 100g de produits, donc logiquement, les apports nutritionnels ne peuvent pas dépasser 100g. De même qu'il ne peut pas avoir de valeurs négatives.

Concernant l'énergie exprimé en KJ, elle ne peut pas dépasser 3700 pour 100g

et les valeurs manquantes par lignes ??

Après ce premier trie, je me concentre sur les valeurs manquantes par lignes et répertorie celles qui n'ont aucunes valeurs renseignées



Le nombre de lignes où il n'y a aucunes informations sur les nutriments est de :26360



Après cette opération, il reste (64706, 11) lignes et colonnes

Imputations

Imputation des valeurs manquantes

Maintenant que notre dataset est nettoyé, je vais pouvoir procéder à l'imputation des valeurs manquantes.

Utilisation de 2 types d'algorithmes différents pour l'imputation :

KNN, Le k-NN est le diminutif de ***k Nearest Neighbors***. C'est un algorithme qui peut servir autant pour la classification que pour la régression. Il est surnommé « nearest neighbors » (plus proches voisins, en français) car le principe de ce modèle consiste en effet à choisir les **k** données les plus proches du point étudié afin d'en prédire sa valeur.

IterativeImputer, Une **imputation initiale** très simple est effectuée : la moyenne est utilisée par défaut. Nous gardons en mémoire les emplacements des valeurs manquantes. Tour à tour — par défaut par ordre décroissant du nombre de valeurs manquantes : Chacune des variables contenant des valeurs manquantes est **exprimée comme fonction des autres variables**, via une régression par exemple ; Les valeurs manquantes sont remplacées via la fonction estimée en a). L'étape 2. constituant ce qui s'appelle un cycle, nous répétons un nombre de cycles (paramètre *max_iter* dans scikit-learn).

Dataset

	code	energy_100g	fat_100g	sugars_100g	salt_100g	saturated-fat_100g	nutrition-score-fr_100g	product_name	pnns_groups_1	nutrition_grade_fr
0	36252	1883.0	20.000000	57.500000	0.096520	12.5000	22.000000	Lion Peanut x2	unknown	e
1	39529	1481.0	4.170000	13.425662	1.154317	5.3712	8.670474	Pack de 2 Twix	unknown	c
2	10187319	1753.0	24.342240	87.700000	0.010000	0.8000	14.000000	Mini Confettis	unknown	d
3	10207260	2406.0	35.469885	50.300000	0.003000	2.9000	14.000000	Praliné Amande Et Noisette	unknown	d
4	40608754	177.0	0.000000	10.400000	0.025400	0.0000	13.000000	Pepsi, Nouveau goût !	Beverages	e
...
64701	9782211109758	1084.0	12.941943	10.500000	0.290000	12.0000	16.000000	Verrine Cheesecake Myrtille	unknown	d
64702	9782401029101	4.0	-5.462125	1.000000	10.000000	1.0000	0.000000	Fiche Brevet	unknown	b
64703	9847548283004	1643.0	2.800000	2.600000	0.680000	0.6000	-4.000000	Tartines craquantes bio au sarrasin	Cereals and potatoes	a
64704	9900000000233	2406.0	35.469885	3.890000	0.100000	3.7300	0.000000	Amandes	unknown	b
64705	99111250	21.0	0.200000	0.500000	0.025400	0.2000	2.000000	Thé vert Earl grey	Beverages	c

64706 rows x 10 columns

Une corrélation a été trouvée entre le nutri-score et le nutri-grade. J'effectue un IterativeImputer pour ces 2 colonnes et un KNN pour le reste.

Le dataset est maintenant prêt pour l'analyse.

Une fonction a été réalisée pour les mises à jour, afin de simplifier le nettoyage.

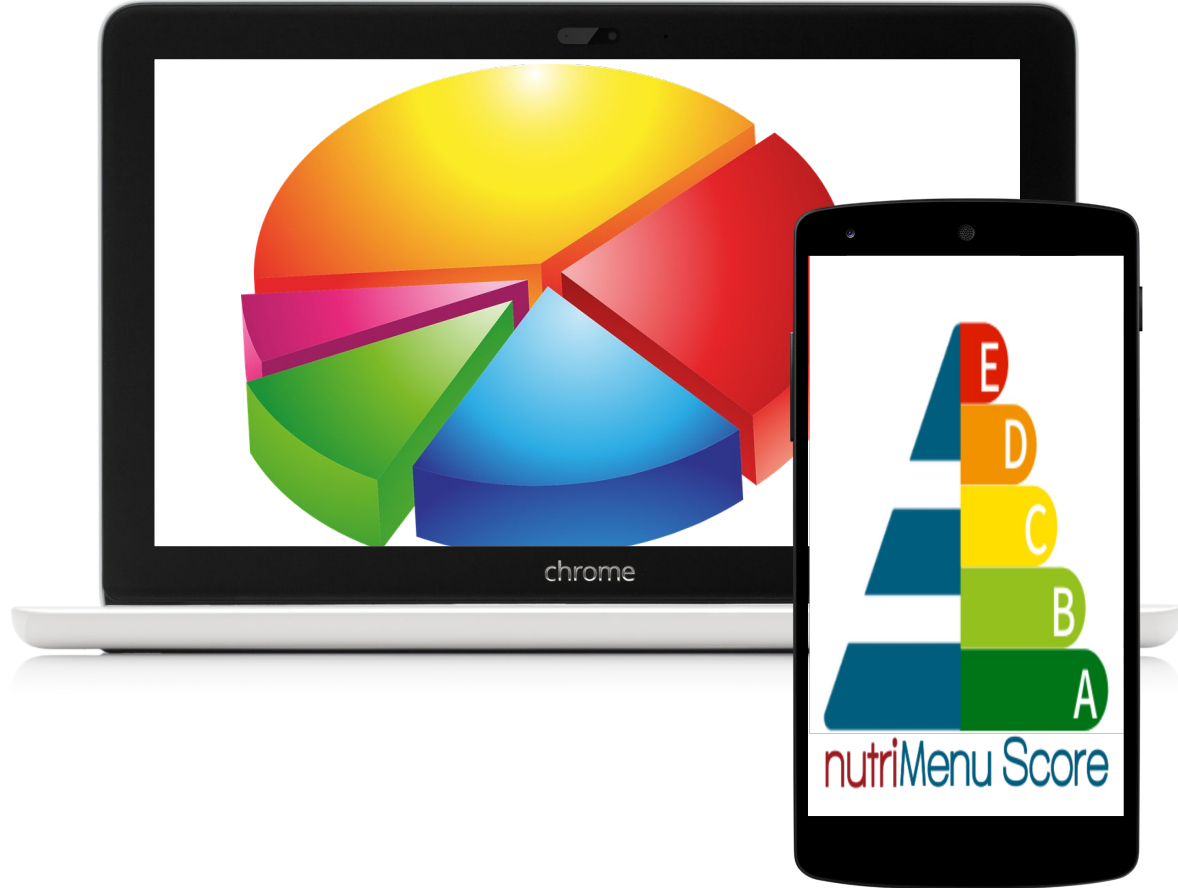


Analyses

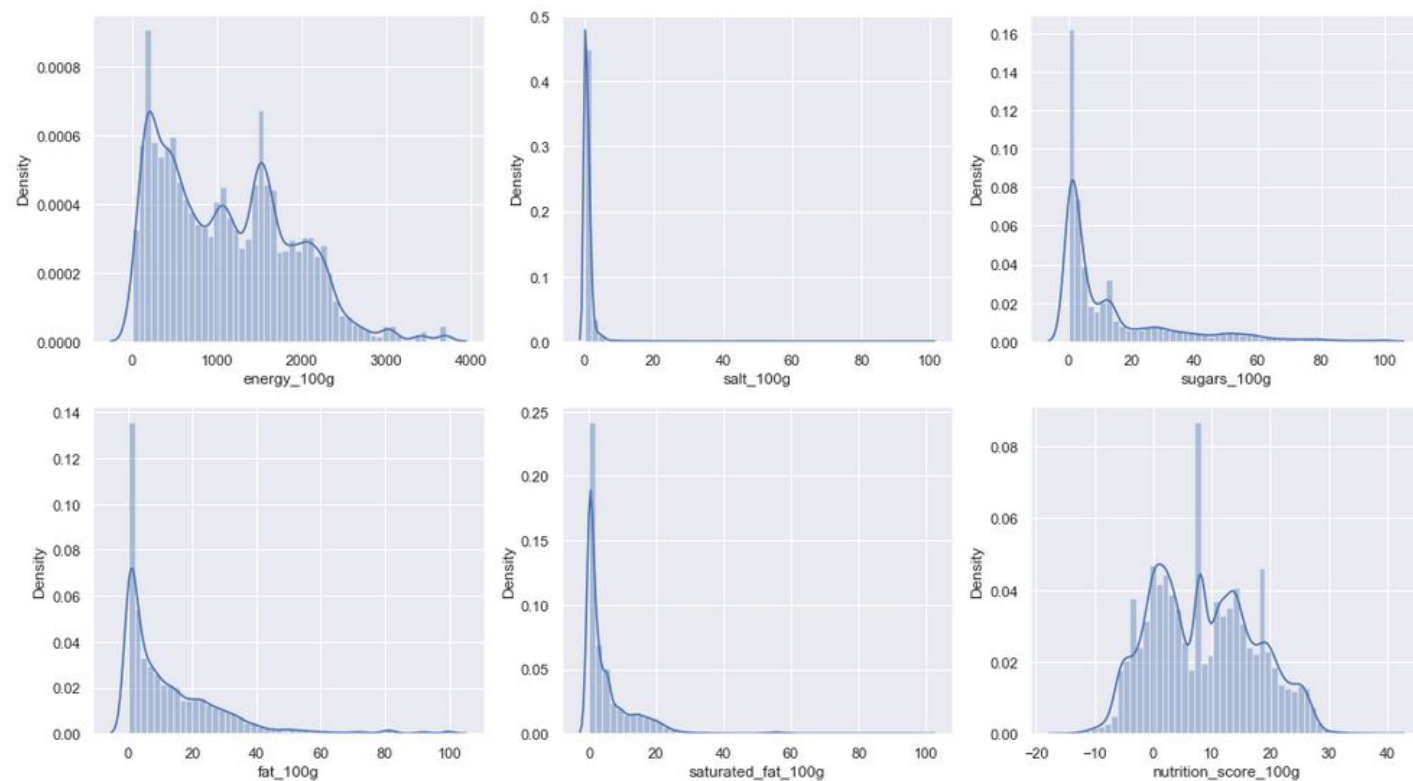
— — —
L'analyse permet de visualiser plusieurs facteurs de nos données afin de mieux les comprendre et de juger de la nécessité d'une application.

A t'on un mauvais nutri-score en fonction des apports alimentaires ?

Les nutri-score non renseigné, sont il systématiquement mauvais ??



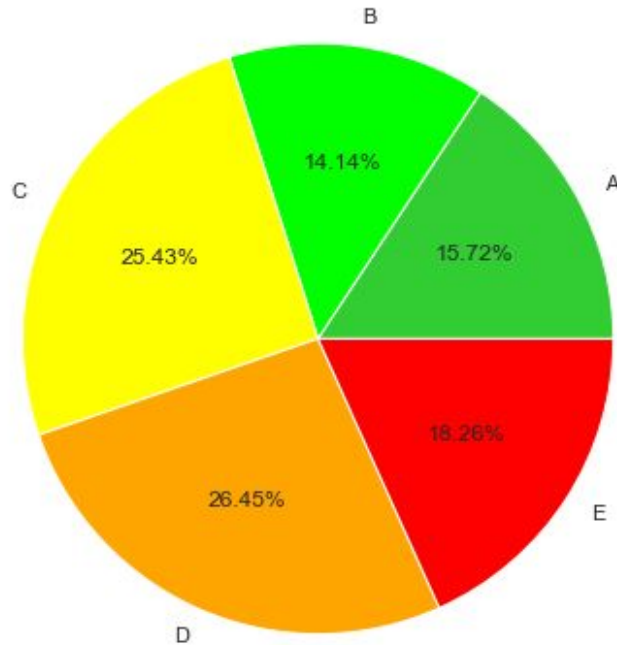
Projection sur un histogramme



Les courbes de ces histogrammes étant les données d'origines (avant les imputations), nous remarquons que les produits qui n'étaient pas renseignés, ont tendance à être des produits de bonnes qualités.

Pourcentage de produits

Pourcentage de produits dans chaque catégorie nutri grade

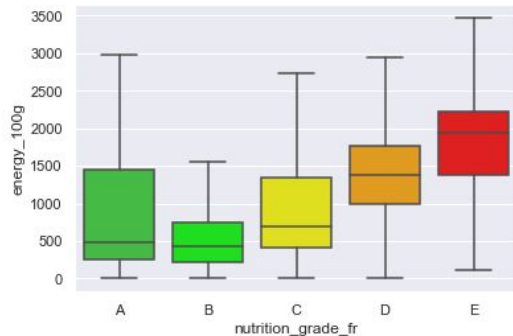


nutrition_grade_fr

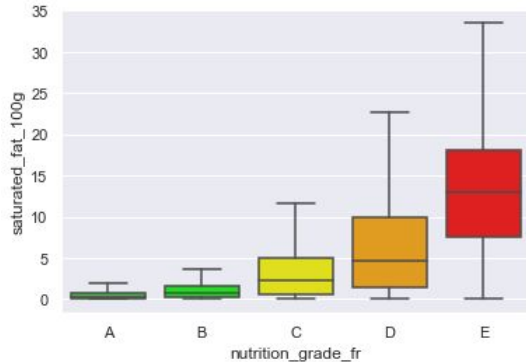
La répartition des produits est un peu plus forte du côté du nutri-grade (D et E), plutôt que (A et B)

Présence d'apport nutritionnel/nutri-score

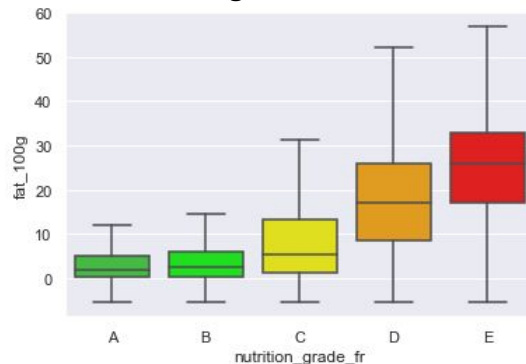
Energy



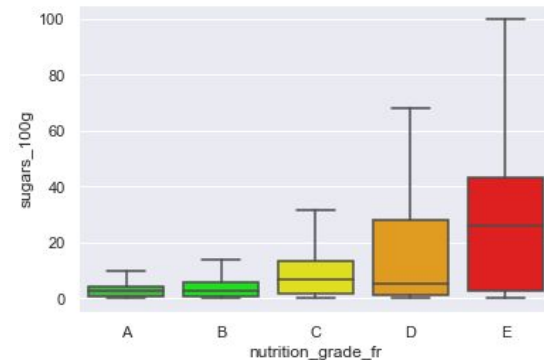
graisses saturées



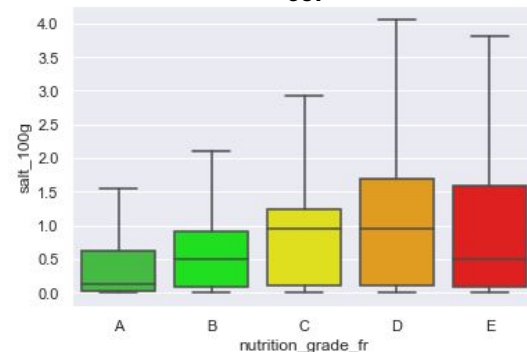
graisses



sucre

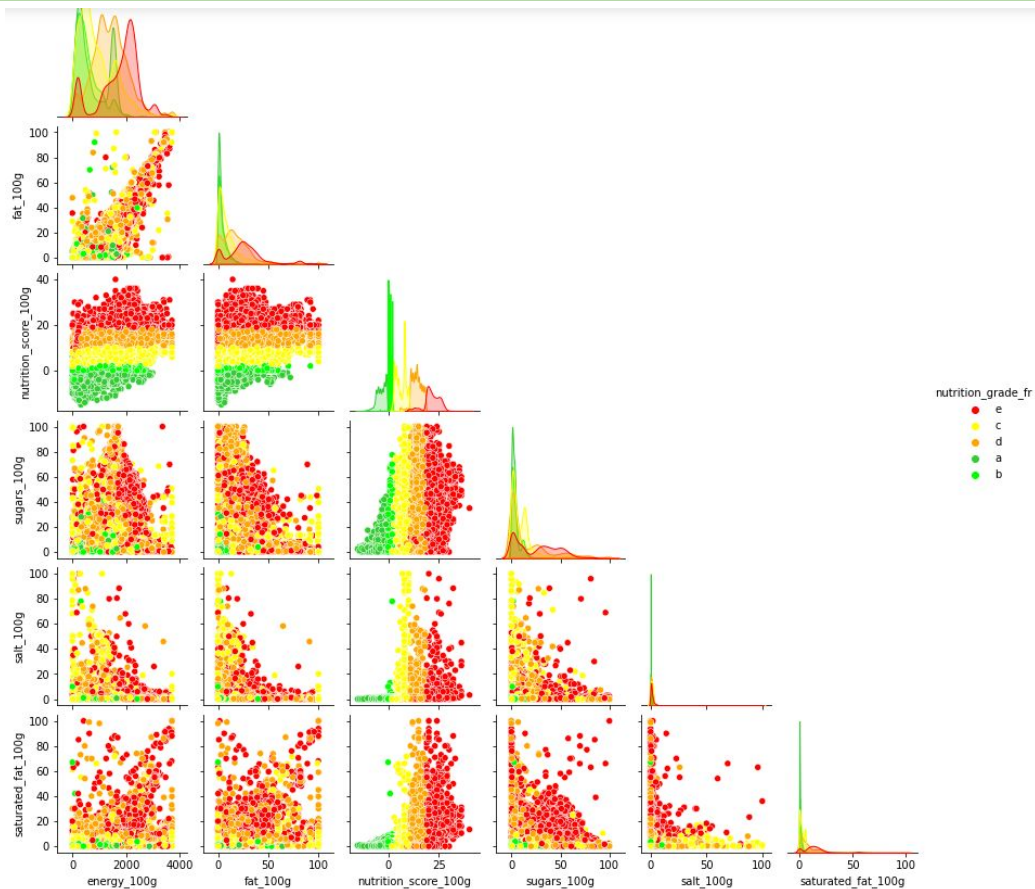


sel



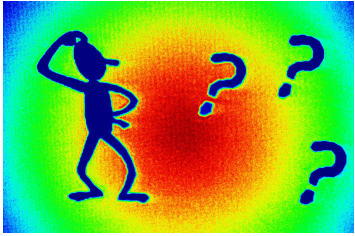
Concernant les apports nutritionnels, nous constatons que plus le nutri-grade est élevé, plus il y a d'apport nutritionnelle

Corrélation entre les apports



Nous pouvons voir clairement que plus le nutri-grade est faible (a, b) moins il y a d'apport.
Corrélation entre l'énergie et les graisses

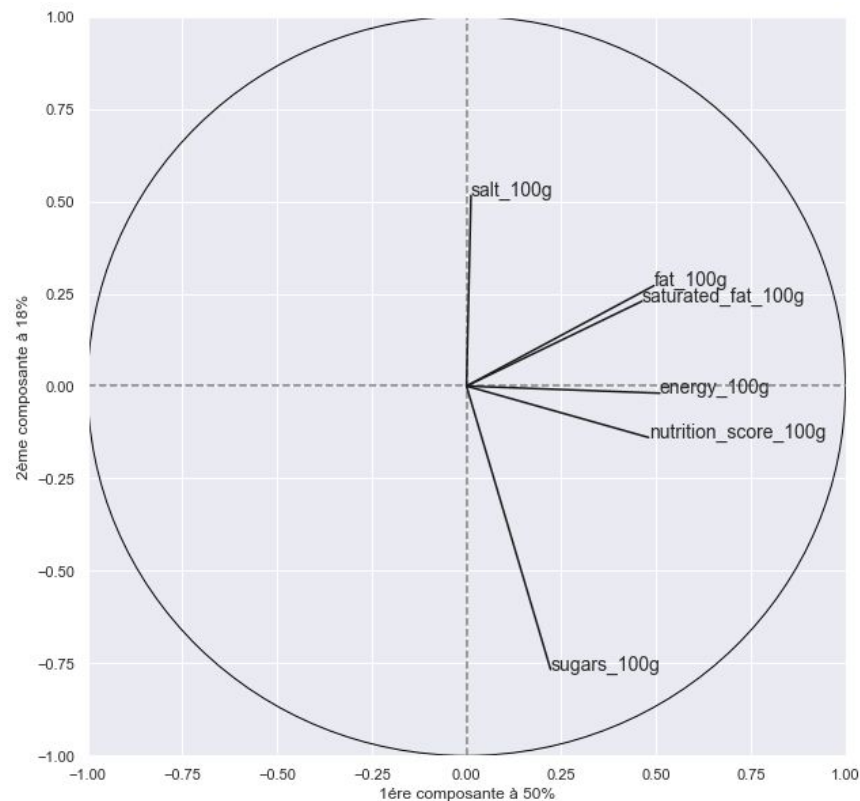
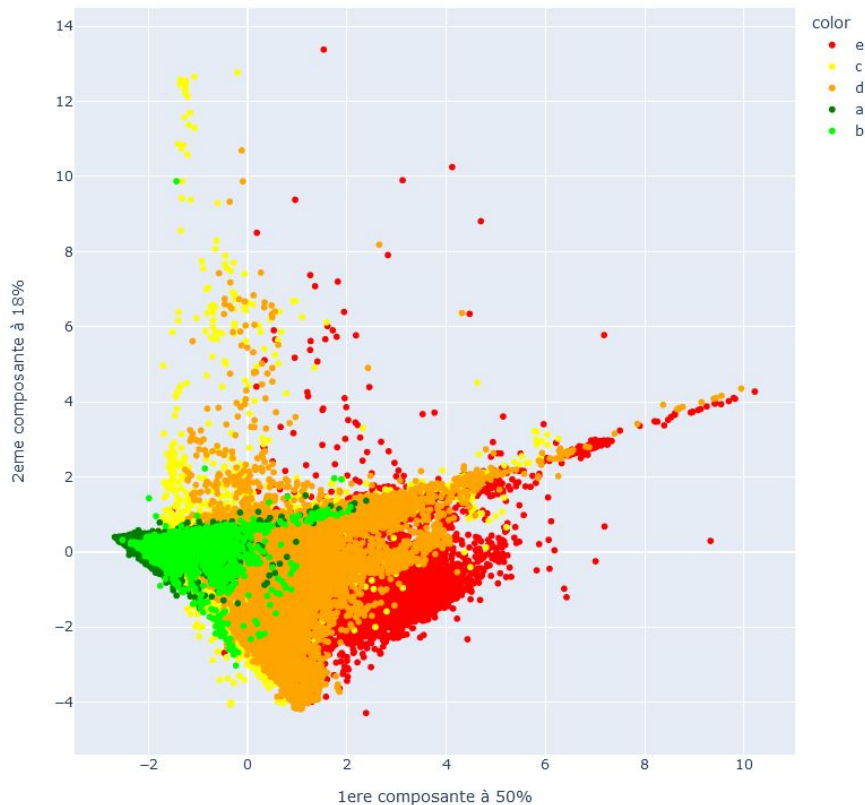
Réalisation d'une ACP



L'analyse en composante principale (ACP), est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites «corrélées») en nouvelles variables décorréliées les unes des autres. Ces nouvelles variables sont nommées « composantes principales ». Elle permet au statisticien de résumer l'information en réduisant le nombre de variables.

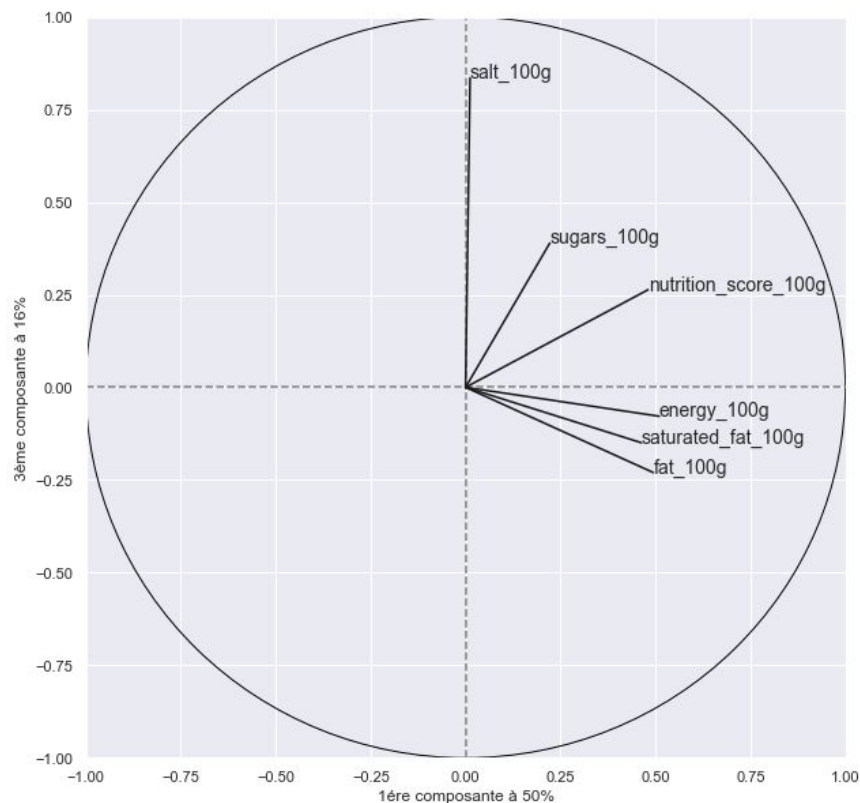
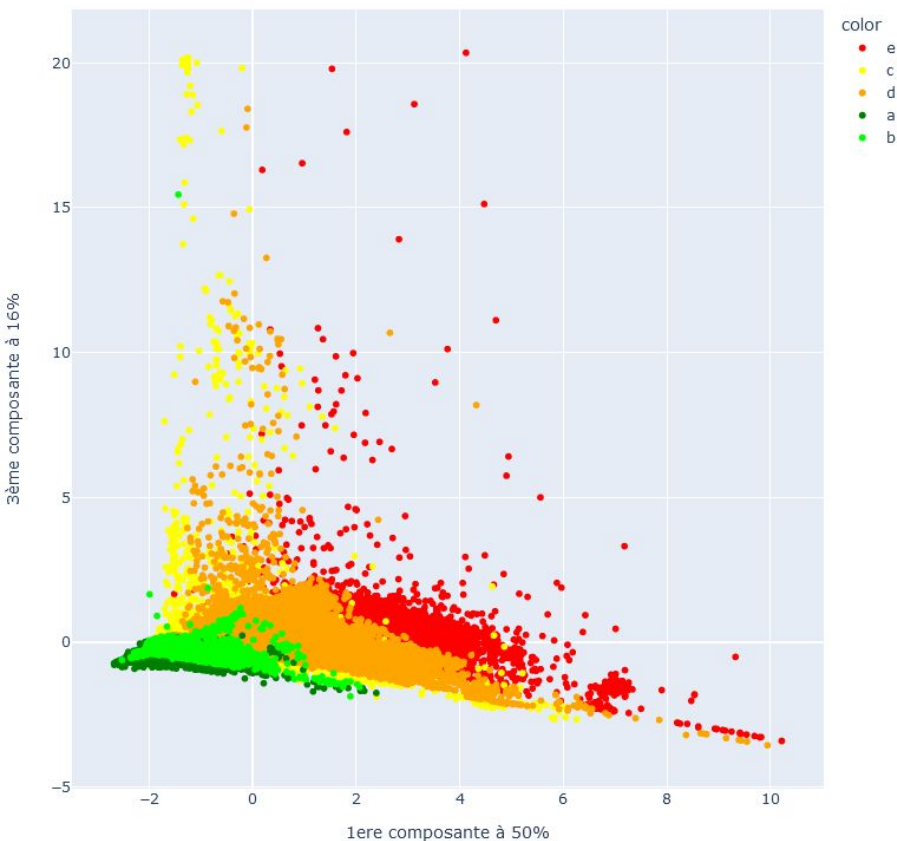
Réalisation d'une ACP

représentation de la 1ère et 2ème composante



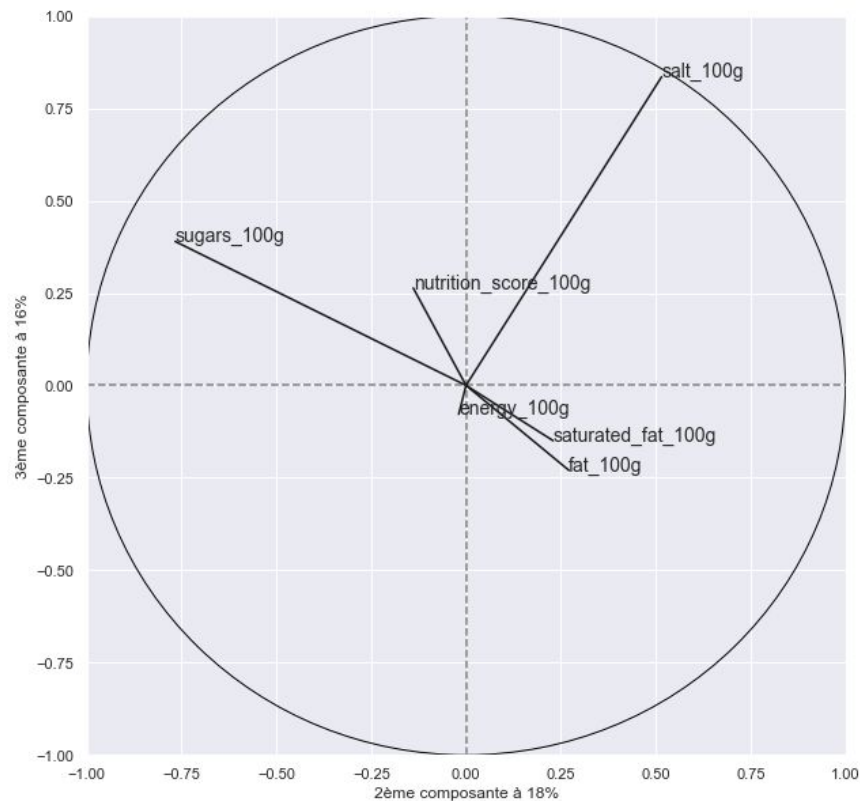
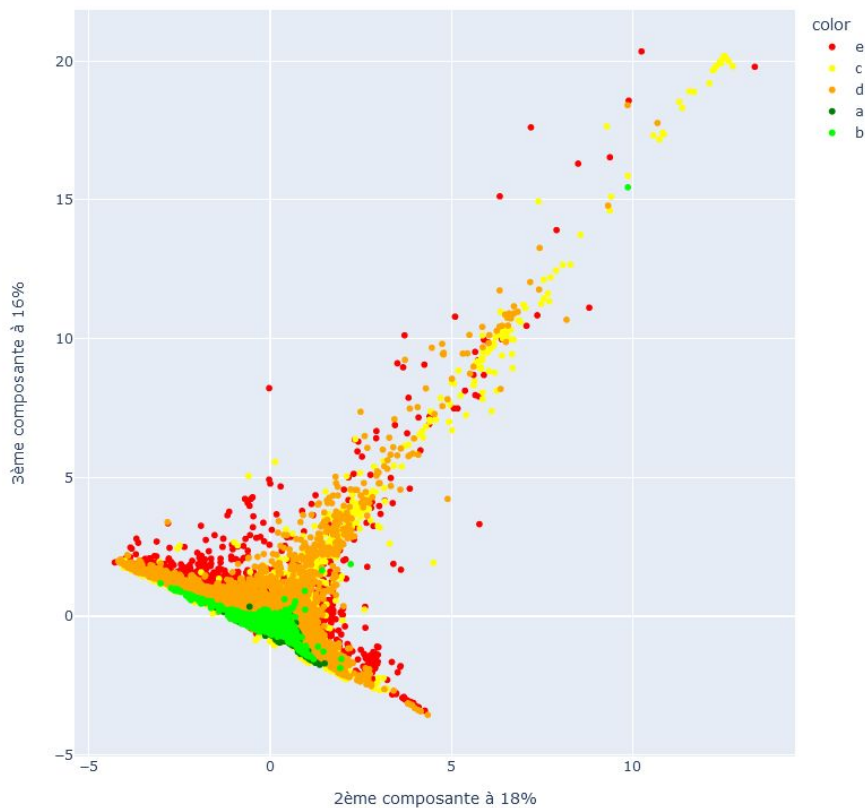
Réalisation d'une ACP

représentation de la 1ère et 3ème composante



Réalisation d'une ACP

représentation de la 2ème et 3ème composante



Réalisation d'une ACP

Conclusion de l'ACP :

- Nous avons tendance à avoir des produits avec beaucoup de graisses, graisses saturées et de sel, dont le nutri-score est d ou e.
- L'énergie et les graisses vont de pairs, donc les graisses apportent l'énergie et apporte un mauvais nutri-score.
- Nous pouvons voir aussi une quantité de produits plus élevé dont le nutri-score est d ou e.
- Avec l'appli, le consommateur sera averti du type de produits qu'il met dans son panier en pouvant vérifier ce qu'il achète.

Conclusion



Il y a une différence significative des produits de mauvaise qualité (D et E) par rapport aux produits de meilleurs qualités (A) et (B) : cela peut s'expliquer par le fait que les entreprises ne souhaitent en général pas afficher le Nutri-Score de leurs produits mal notés. En effet Le Nutri-Score est devenu une obligation au 1er janvier 2021. Toutefois « Les annonceurs et les promoteurs peuvent déroger à cette obligation sous réserve du versement d'une contribution dont le produit est affecté à l'Agence nationale de santé publique ». Ce versement est dû dès lors que les fabricants et distributeurs du secteur alimentaire décident de ne pas faire figurer le Nutri-Score lors de la diffusion de leurs messages publicitaires et sur leurs emballages produits. L'application permettra donc facilement de fournir à l'utilisateur la confirmation que oui ou non, le produit non noté est de mauvaise qualité nutritionnelle suivant le cahier des charges "Santé Publique France".