

# Préparez des données pour un organisme de santé publique



# Sommaire

Contexte

1

Présentation  
du jeu de  
données

2

3

Nettoyage de  
la base de  
données

5

4

Analyse  
univariée

6

Conclusion



# 1. Contexte



## Notre mission

- **Objectif :** Améliorer la base de données Open Food Facts pour faciliter l'ajout de nouveaux produits.
- **Défis actuels :** Nombre élevé de champs à remplir entraînant des oubli et des erreurs.
- **Mission :** Explorer des solutions pour fournir des suggestions ou une fonction d'auto-complétion lors de la saisie des données.
- **Bénéficiaires :** Particuliers et organisations souhaitant contribuer à la base de données Open Food Facts.
- **But :** Simplifier et améliorer le processus de collecte d'informations.

## 2. Le jeu de données

*Taille du jeu de données :*

320 772 lignes et 162 colonnes.

*Lignes :*

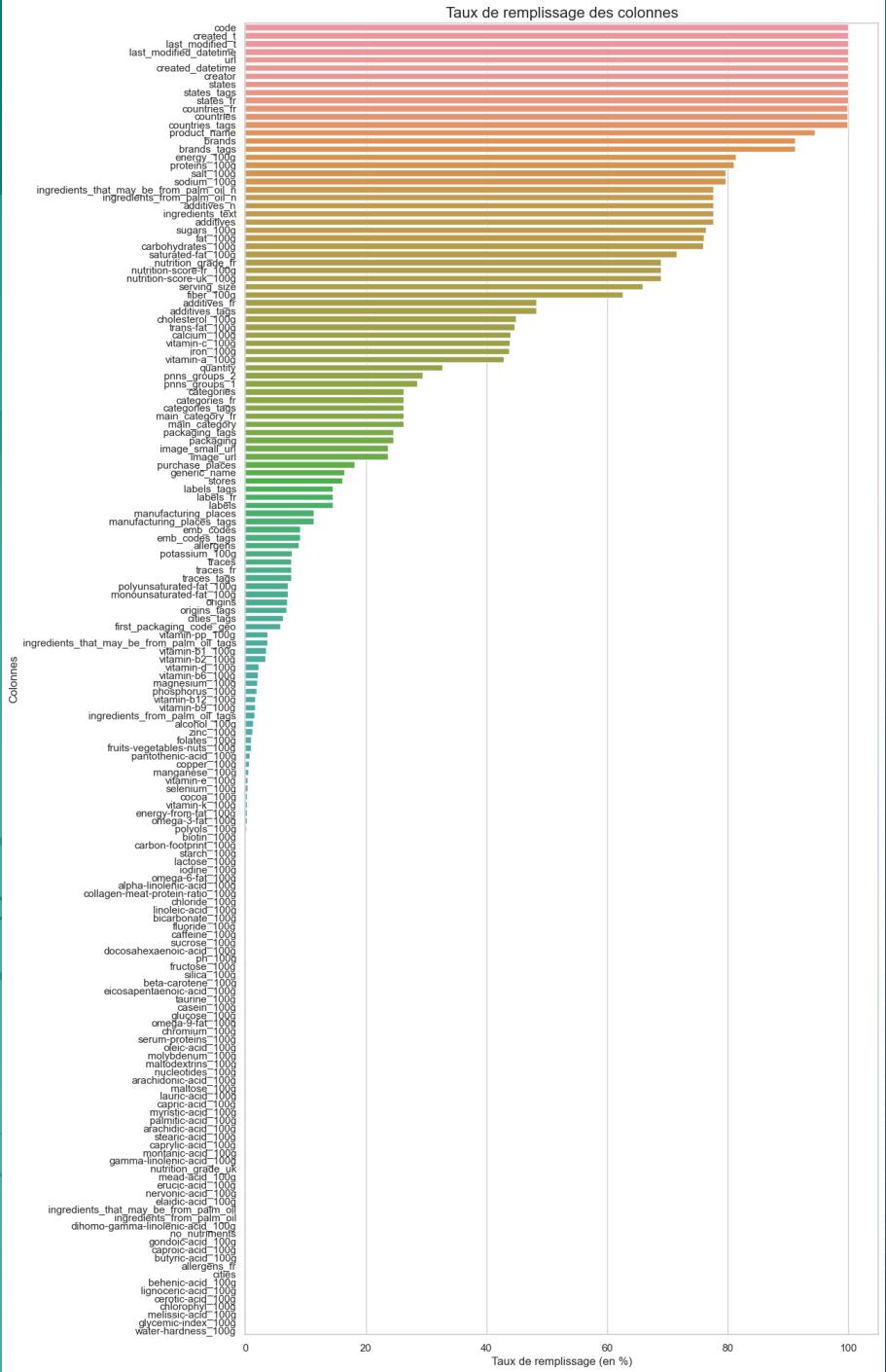
Chaque entrée représente un produit alimentaire unique.

*Variables :*

Composé de 106 variables quantitatives et 56 variables qualitatives.

*Problème majeur :*

Présence de données manquantes, dépassant 75% dans de nombreuses colonnes.



### 3. Nettoyage de la base de données

Suppression de colonnes et de lignes

Suppression des colonnes :

- Élimination des colonnes avec plus de 75% de valeurs manquantes.
- Élimination des colonnes similaires (>70% de similarité) grâce à la distance de Levenshtein.
- Suppression des colonnes non pertinentes pour la mission de Santé Publique France.

Suppression des lignes :

- Suppression des produits non identifiés (sans code).
- Exclusion des entrées sans noms de produits.
- Suppression des produits ne contenant aucune données nutritionnelles.

Résultat : Réduction de la taille du jeu de données de 320,772 à 259,539 lignes et de 162 à 28 colonnes.



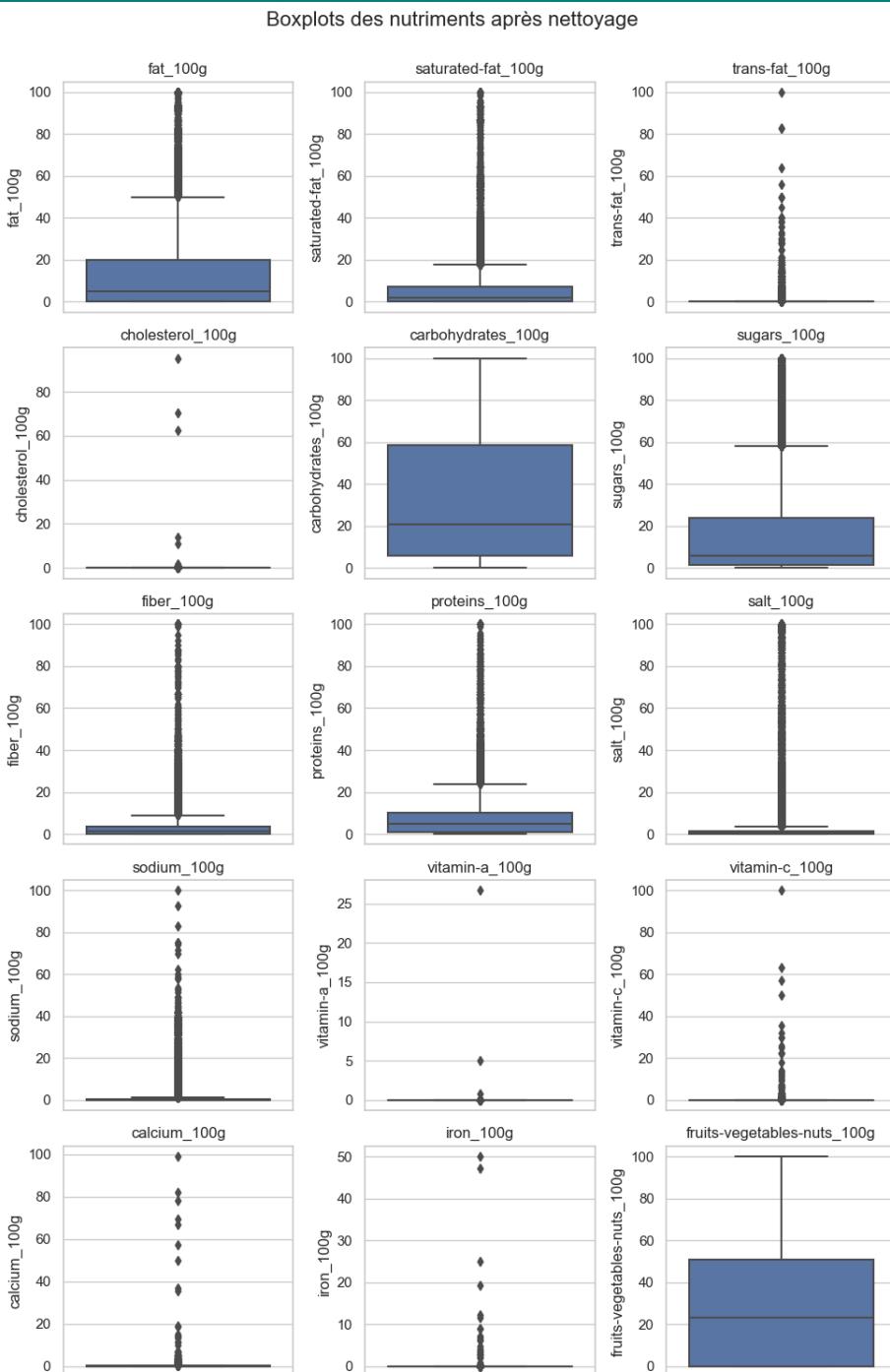
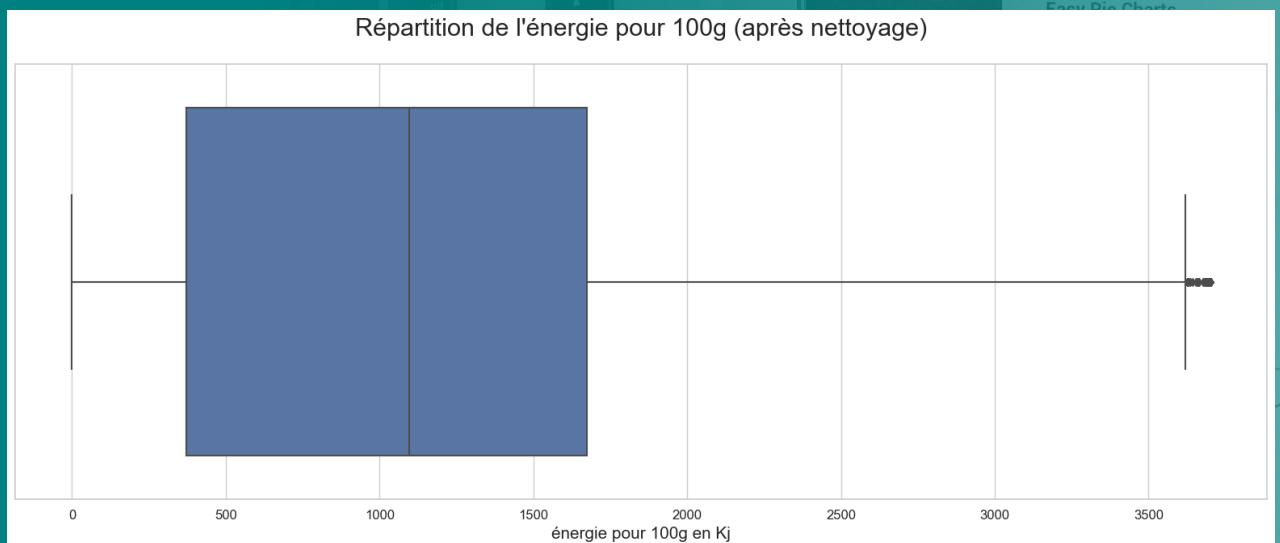
Dimension du jeu de données : (259539, 28)

B	class 'pandas.core.frame.DataFrame'>		
	RangeIndex: 259539 entries, 0 to 259538		
	Data columns (total 28 columns):		
#	Column	Non-Null Count	Dtype
---	---	-----	-----
0	code	259539	non-null
1	product_name	259539	non-null
2	brands	256167	non-null
3	categories_fr	64109	non-null
4	countries_fr	259474	non-null
5	additives_n	234027	non-null
6	additives_fr	149577	non-null
7	ingredients_from_palm_oil_n	234027	non-null
8	nutrition_grade_fr	218463	non-null
9	pnns_groups_2	68369	non-null
10	main_category_fr	64109	non-null
11	energy_100g	257773	non-null
12	fat_100g	240584	non-null
13	saturated-fat_100g	226641	non-null
14	trans-fat_100g	143159	non-null
15	cholesterol_100g	143950	non-null
16	carbohydrates_100g	240276	non-null
17	sugars_100g	241910	non-null
18	fiber_100g	198587	non-null
19	proteins_100g	256605	non-null
20	salt_100g	252527	non-null
21	sodium_100g	252488	non-null
22	vitamin-a_100g	137398	non-null
23	vitamin-c_100g	140655	non-null
24	calcium_100g	140837	non-null
25	iron_100g	140305	non-null
26	fruits-vegetables-nuts_100g	3032	non-null
27	nutrition-score-fr_100g	218463	non-null

### 3. Nettoyage de la base de données

Recherche de valeurs aberrantes ou atypiques

- Valeurs aberrantes dans les valeurs nutritionnelles car ces valeurs doivent être comprises entre 0 et 100 g pour 100 g de produit.
- Valeurs aberrantes également pour l'énergie car le minimum d'énergie en kJ est de 0 pour l'eau et 3700 pour les aliments les plus caloriques (huiles).
- Nettoyage des Valeurs Aberrantes : Remplacement par des valeurs manquantes que nous imputerons après.



### 3. Nettoyage de la base de données

#### Gestion des valeurs manquantes .1/2

##### Objectif

Gérer les valeurs manquantes de manière appropriée pour assurer l'intégrité des données.

##### Remplacement des valeurs manquantes non estimables

Substitution des valeurs manquantes par un texte explicite.

Exemple : Remplacer les additifs manquants par "sans additifs".

##### Comparaison de Méthodes d'Imputation pour les autres valeurs manquantes

Trois méthodes testées : moyenne, médiane, KNNImputer.

##### Utilisation du KNNImputer

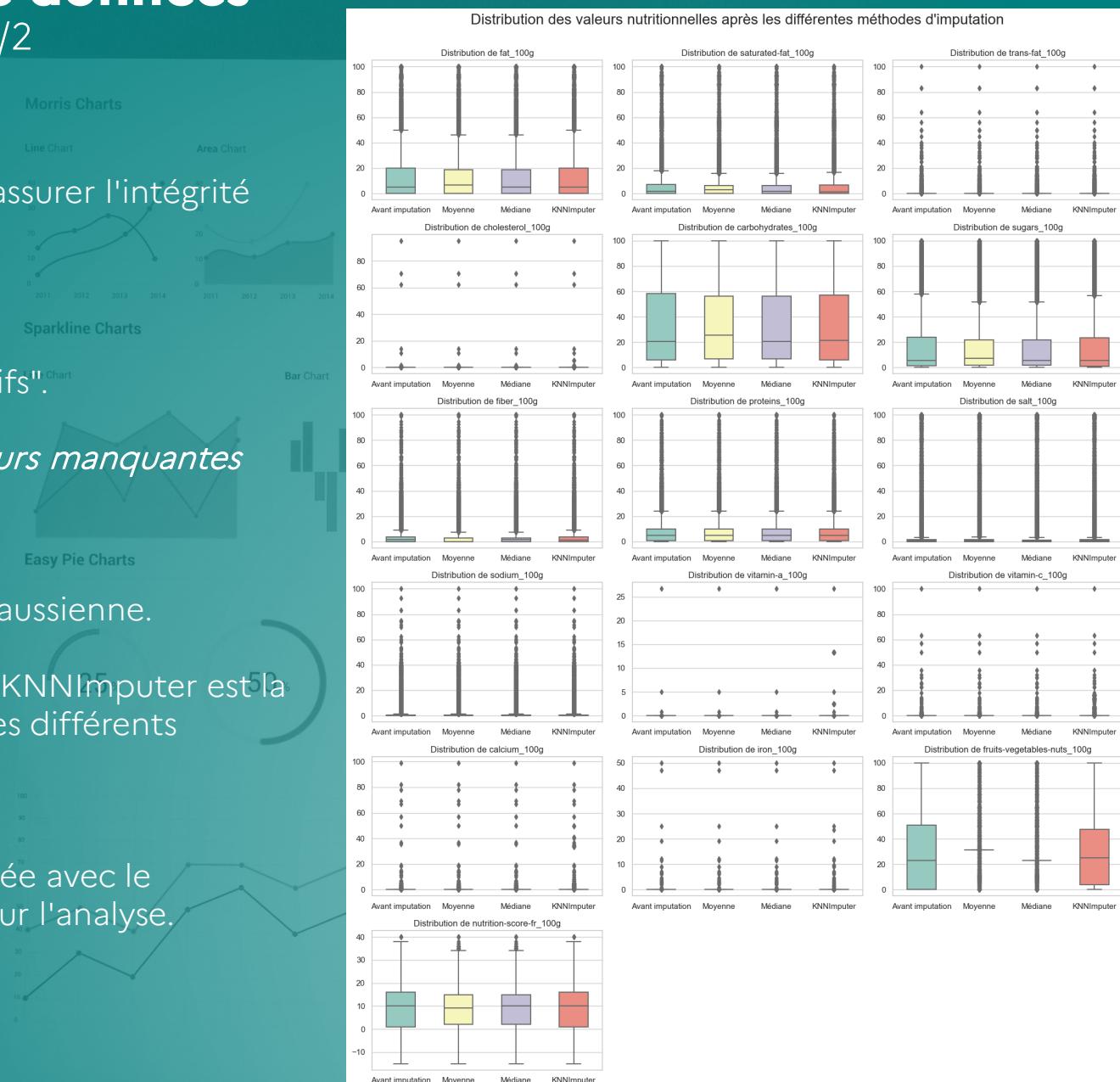
Les distributions des nutriments ne suivent pas une forme gaussienne.

L'imputation par la moyenne serait inadaptée.

La médiane est moins sensible aux valeurs extrêmes mais le KNNImputer est la meilleure option pour son adaptation aux variations entre les différents aliments.

##### Résultat : Imputation des Valeurs Manquantes

Les valeurs manquantes ont été gérées de manière appropriée avec le KNNImputer, garantissant que les données soient prêtes pour l'analyse.



### 3. Nettoyage de la base de données

#### Gestion des valeurs manquantes .2/2

##### Énergie Manquante

Nous avons estimé l'énergie des produits en Kj à partir des protéines, des glucides et des matières grasses, conformément aux coefficients d'Atwater. Les valeurs manquantes ont été remplacées par les résultats de ce calcul.

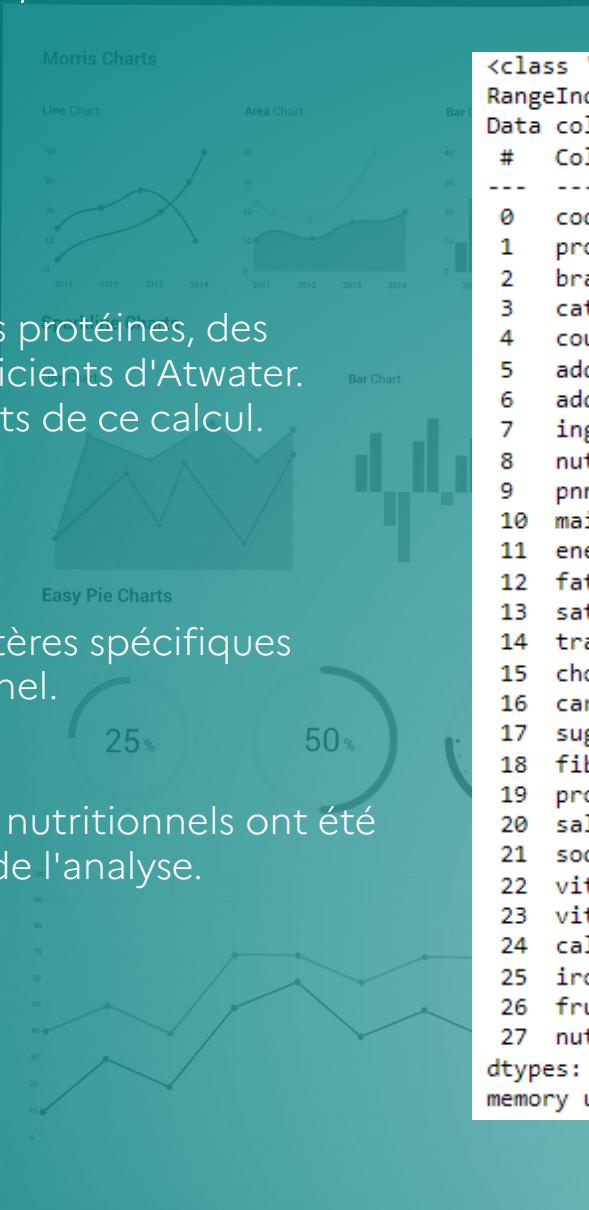
##### Grades Nutritionnels

Nous avons assigné des grades nutritionnels aux produits. Les eaux ont été automatiquement classées en grade "A".

Pour les boissons et les aliments, nous avons utilisé des critères spécifiques pour déterminer les grades en fonction du score nutritionnel.

##### Résultat

Toutes les valeurs manquantes pour l'énergie et les grades nutritionnels ont été comblées, garantissant la cohérence des données en vue de l'analyse.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 259539 entries, 0 to 259538
Data columns (total 28 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   code             259539 non-null  object  
 1   product_name     259539 non-null  object  
 2   brands            259539 non-null  object  
 3   categories_fr    259539 non-null  object  
 4   countries_fr     259539 non-null  object  
 5   additives_n       259539 non-null  float64 
 6   additives_fr      259539 non-null  object  
 7   ingredients_from_palm_oil_n  259539 non-null  float64 
 8   nutrition_grade_fr 259539 non-null  object  
 9   pnns_groups_2     259539 non-null  object  
 10  main_category_fr 259539 non-null  object  
 11  energy_100g       259539 non-null  float64 
 12  fat_100g          259539 non-null  float64 
 13  saturated-fat_100g 259539 non-null  float64 
 14  trans-fat_100g    259539 non-null  float64 
 15  cholesterol_100g 259539 non-null  float64 
 16  carbohydrates_100g 259539 non-null  float64 
 17  sugars_100g        259539 non-null  float64 
 18  fiber_100g         259539 non-null  float64 
 19  proteins_100g      259539 non-null  float64 
 20  salt_100g           259539 non-null  float64 
 21  sodium_100g         259539 non-null  float64 
 22  vitamin-a_100g      259539 non-null  float64 
 23  vitamin-c_100g      259539 non-null  float64 
 24  calcium_100g        259539 non-null  float64 
 25  iron_100g           259539 non-null  float64 
 26  fruits-vegetables-nuts_100g 259539 non-null  float64 
 27  nutrition-score-fr_100g   259539 non-null  float64 
dtypes: float64(19), object(9)
memory usage: 55.4+ MB
```

# 4. Analyse univariée

## Distribution des variables

### *Caractéristiques des Distributions*

Les distributions ne suivent pas une loi normale.

Les valeurs modales sont généralement proches de zéro, sauf pour l'énergie et le nutrition-score où elles sont bimodales.

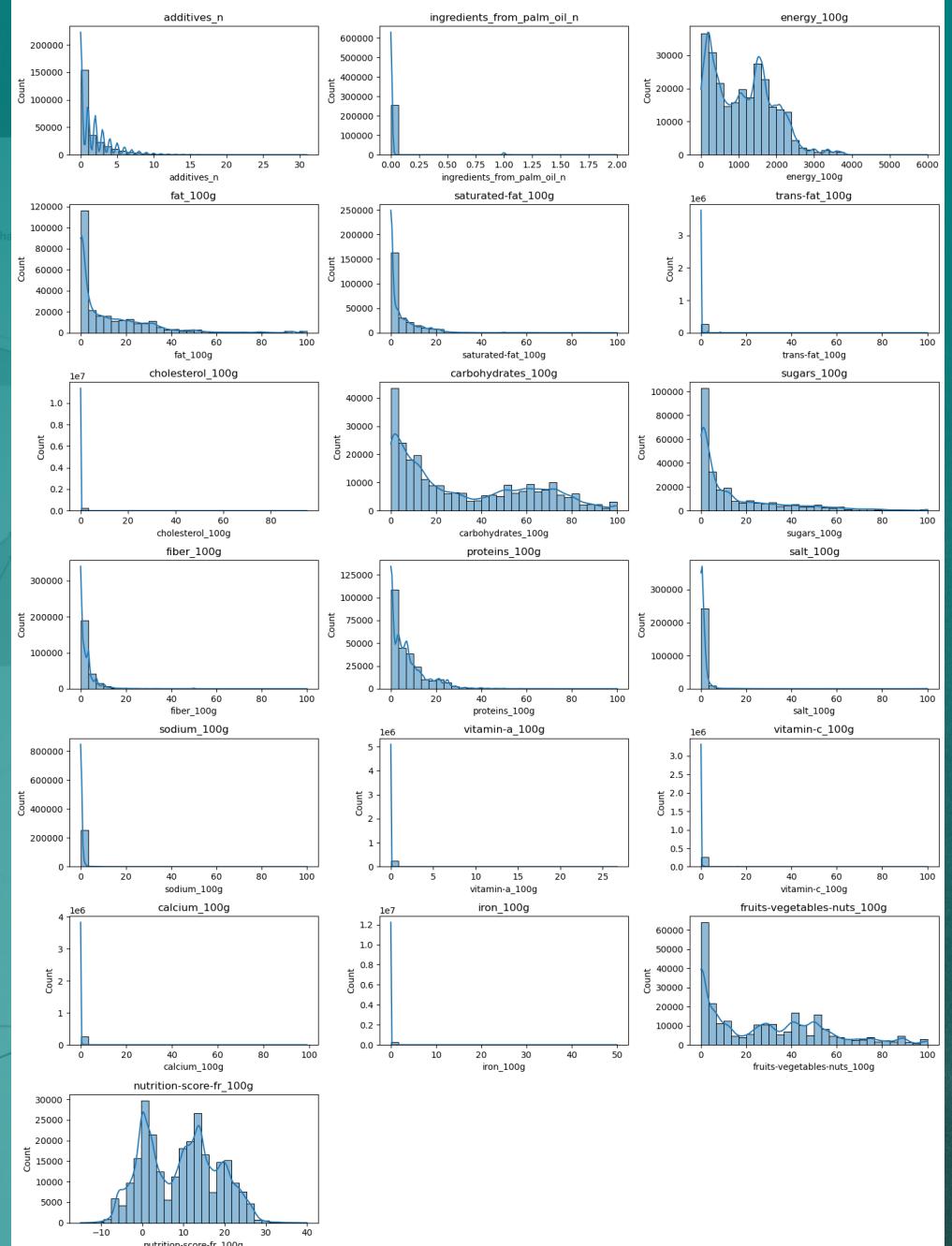
Les distributions sont étalées vers la droite, indiquant des valeurs élevées.

### *Relation entre Mode, Médiane et Moyenne*

Dans une distribution étalée vers la droite :

- Le mode est inférieur à la médiane.
- La médiane est inférieure à la moyenne.

Distribution des variables quantitatives



## 4. Analyse univariée

## Répartition des grades Nutri-score et noms de produits

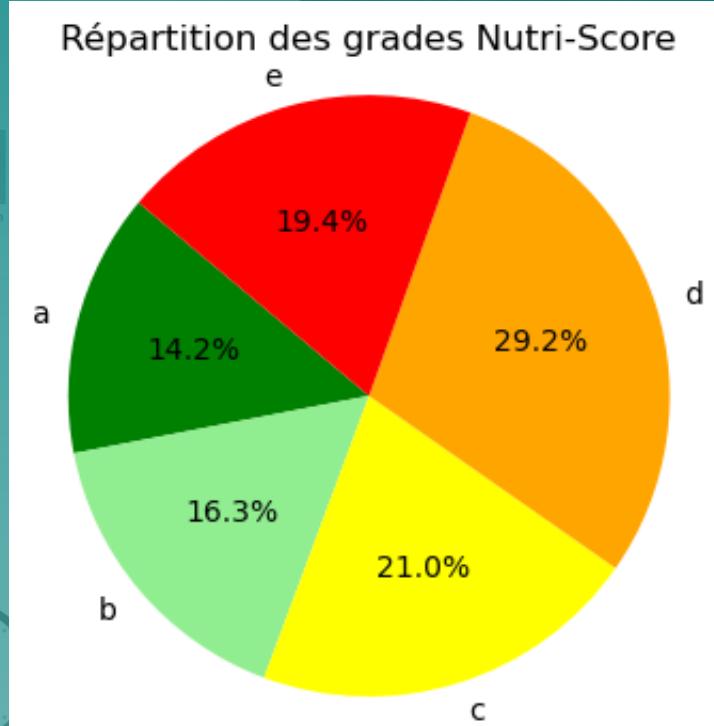
## Répartition des Grades Nutri-Score

## Analyse de la répartition des produits en fonction de leur Nutri-Score.

Implication : Présence dominante de produits potentiellement moins sains.

## *Nuage de Mots des Noms de Produits*

Prédominance de produits sucrés tels que le chocolat et la crème glacée.



# 5. Analyse multivariée

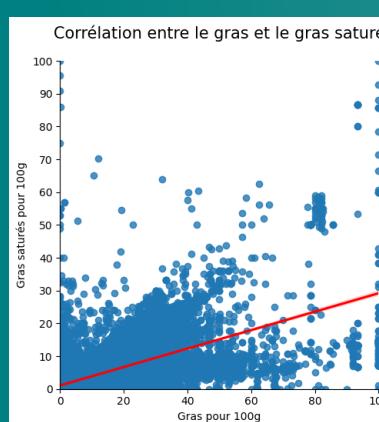
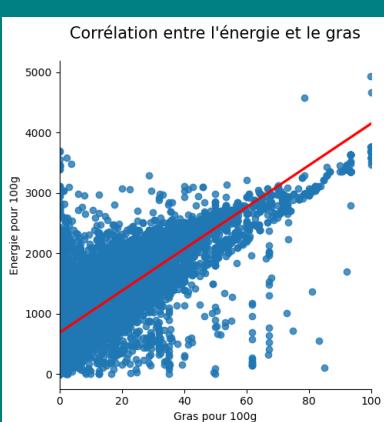
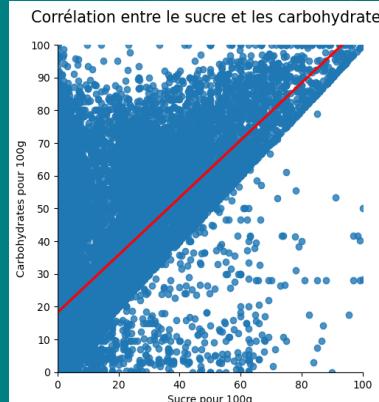
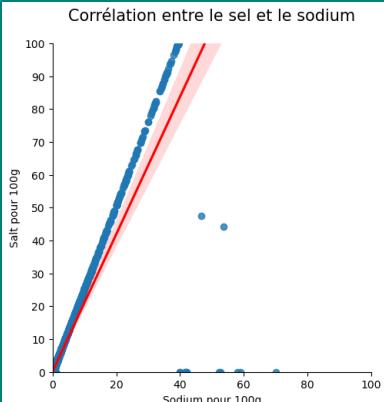
## Matrice des corrélations

### *Introduction à l'Analyse Multivariée*

Une analyse bivariée permet de détecter les corrélations entre les variables.

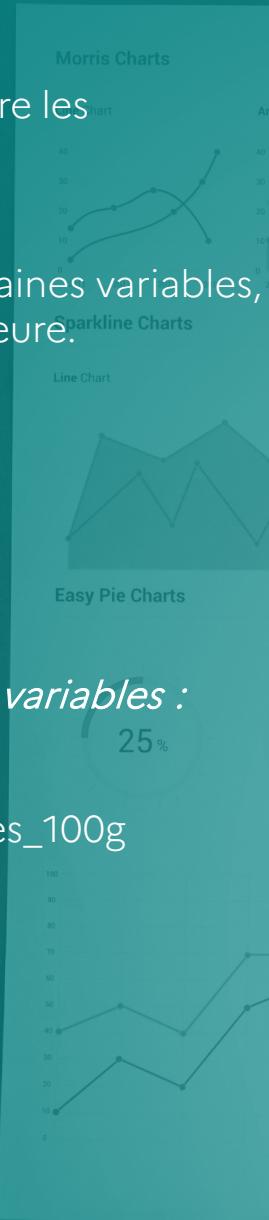
### *Conséquences pour l'Analyse Multivariée*

La détection de corrélations nous permet de supprimer certaines variables, simplifiant l'analyse en composantes principales (ACP) ultérieure.

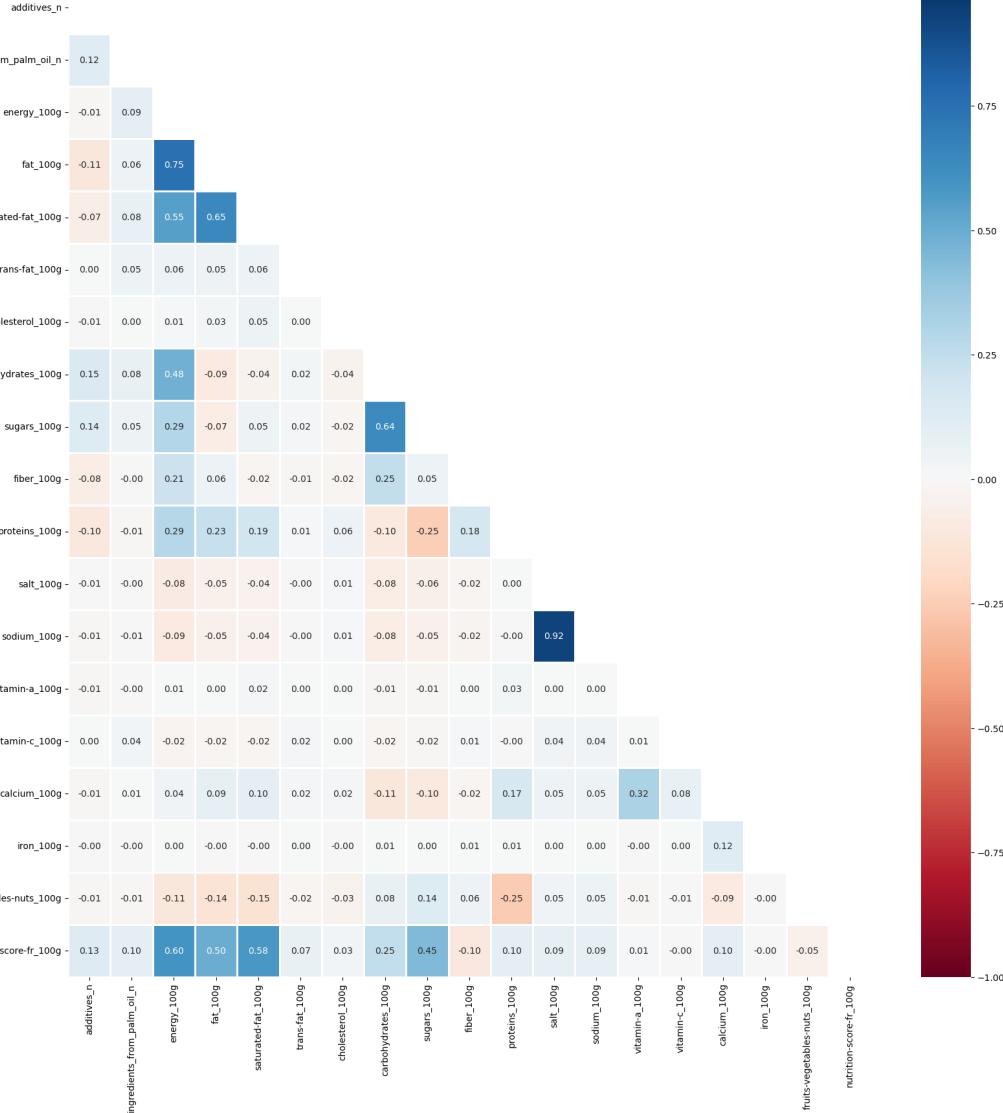


### *Suppression des variables :*

- Sodium\_100g
- Fat\_100g
- Carbohydrates\_100g



Matrice de corrélation de Pearson

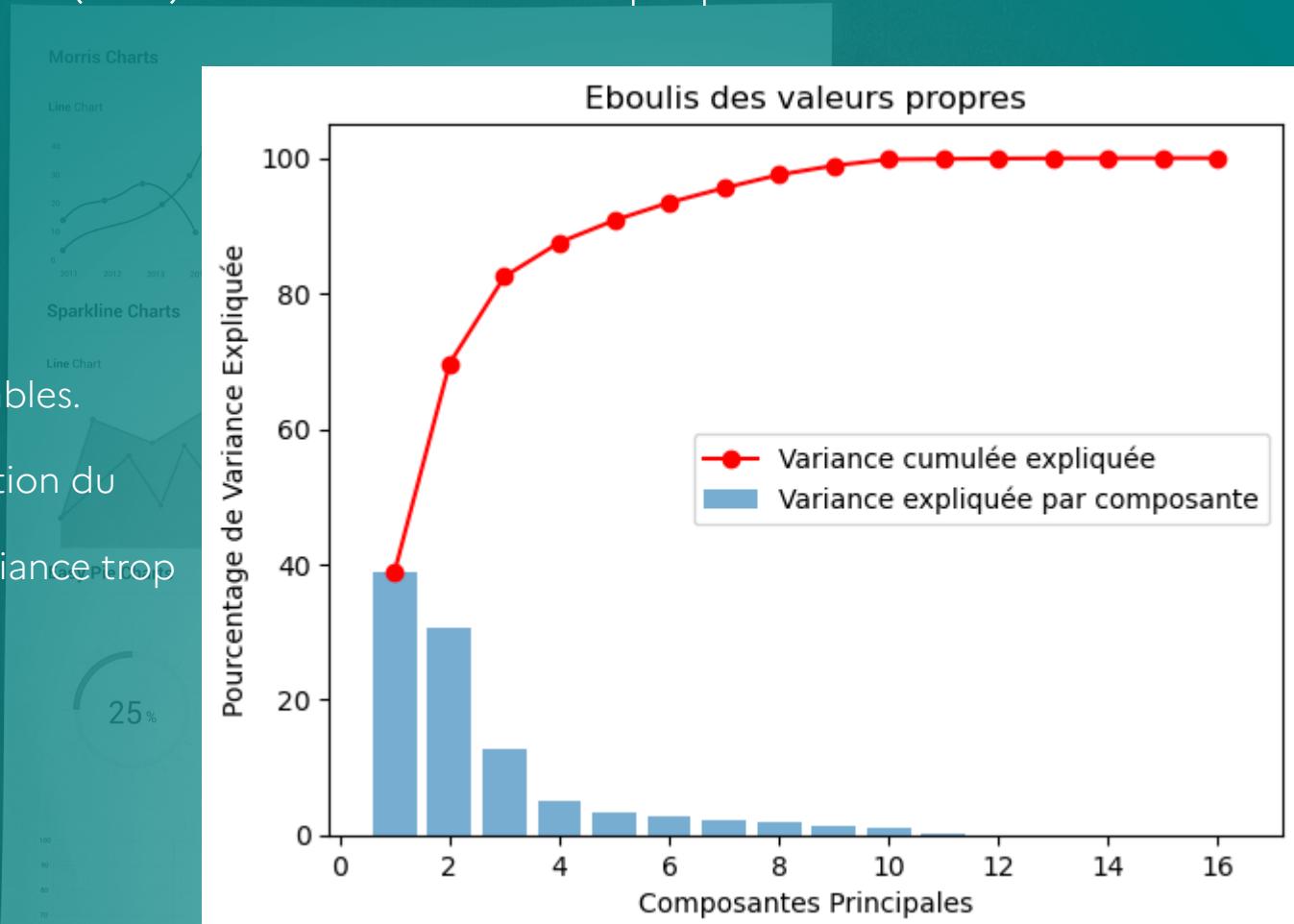


## 5. Analyse multivariée

Analyse en composantes principales (ACP) – Eboulis des valeurs propres

Nous avons réalisé une ACP afin de réduire le nombre de variables.

- Les 4 premières composantes captent 87.57% de l'information du Dataset.
- Les autres composantes expliquent un pourcentage de variance trop faible pour être significatif.

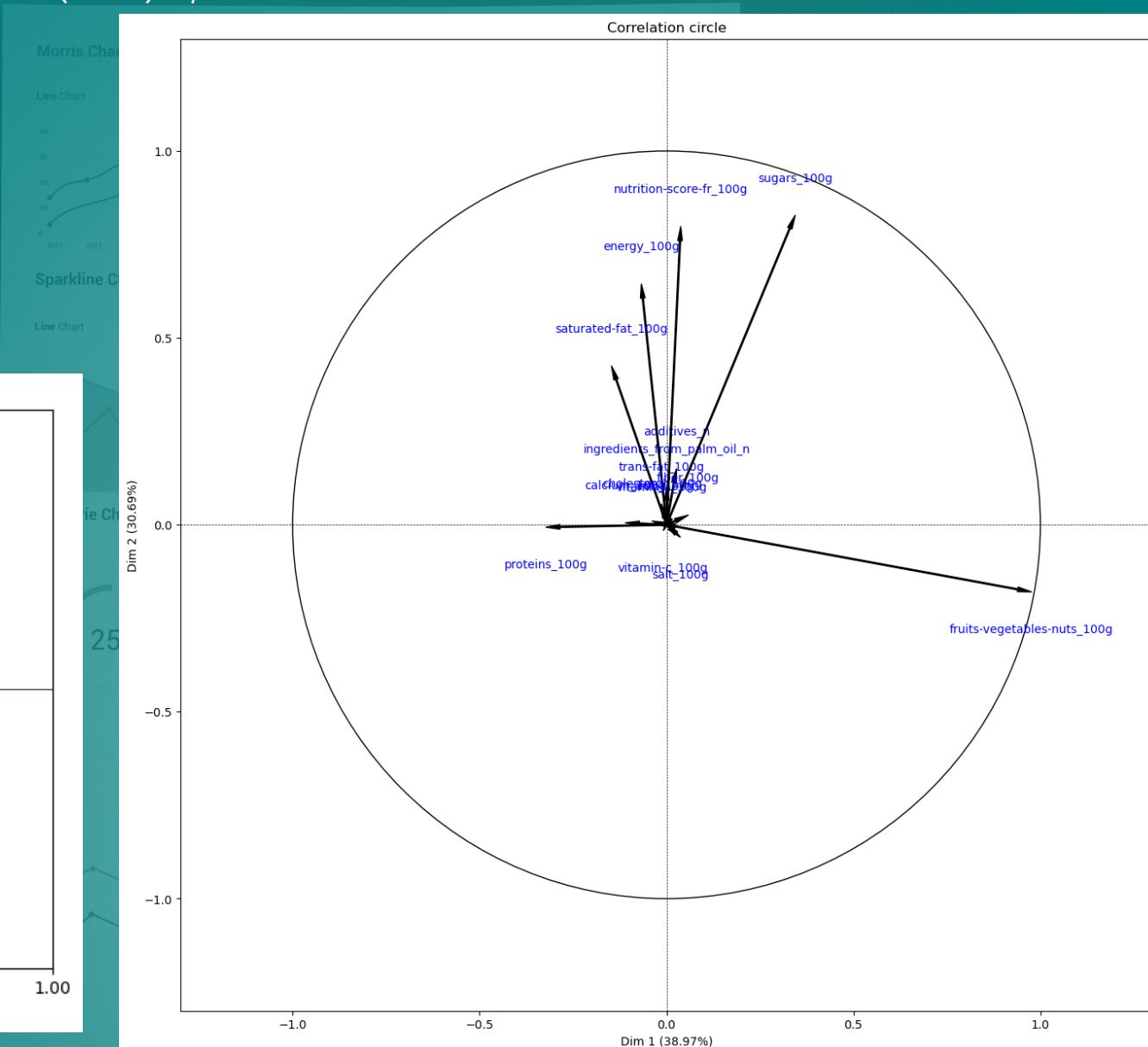
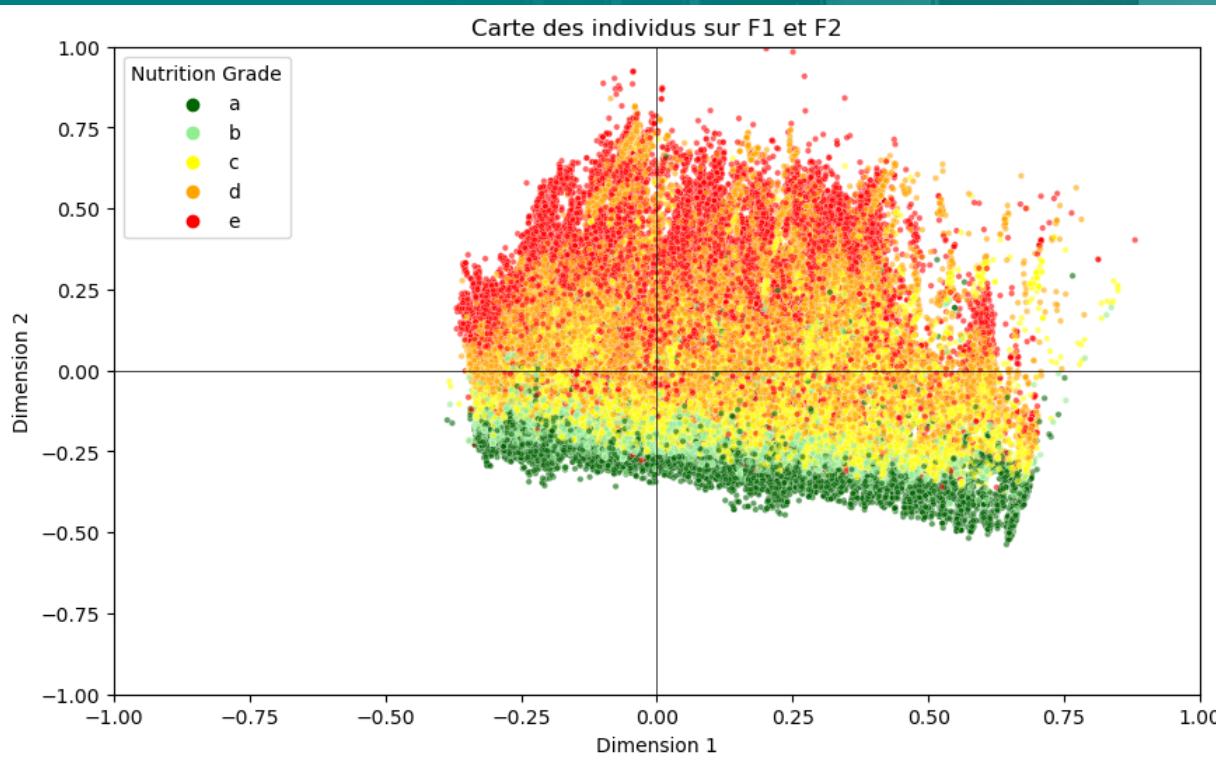


## 5. Analyse multivariée

Analyse en composantes principales (ACP) .1/2

### *Interprétation des composantes F1 et F2*

- F1 : corrélation positive avec les produits à base de fruits et légumes.
- F2 : corrélation positive avec les produits sucrés, énergétiques, gras et avec un haut score nutritionnel.

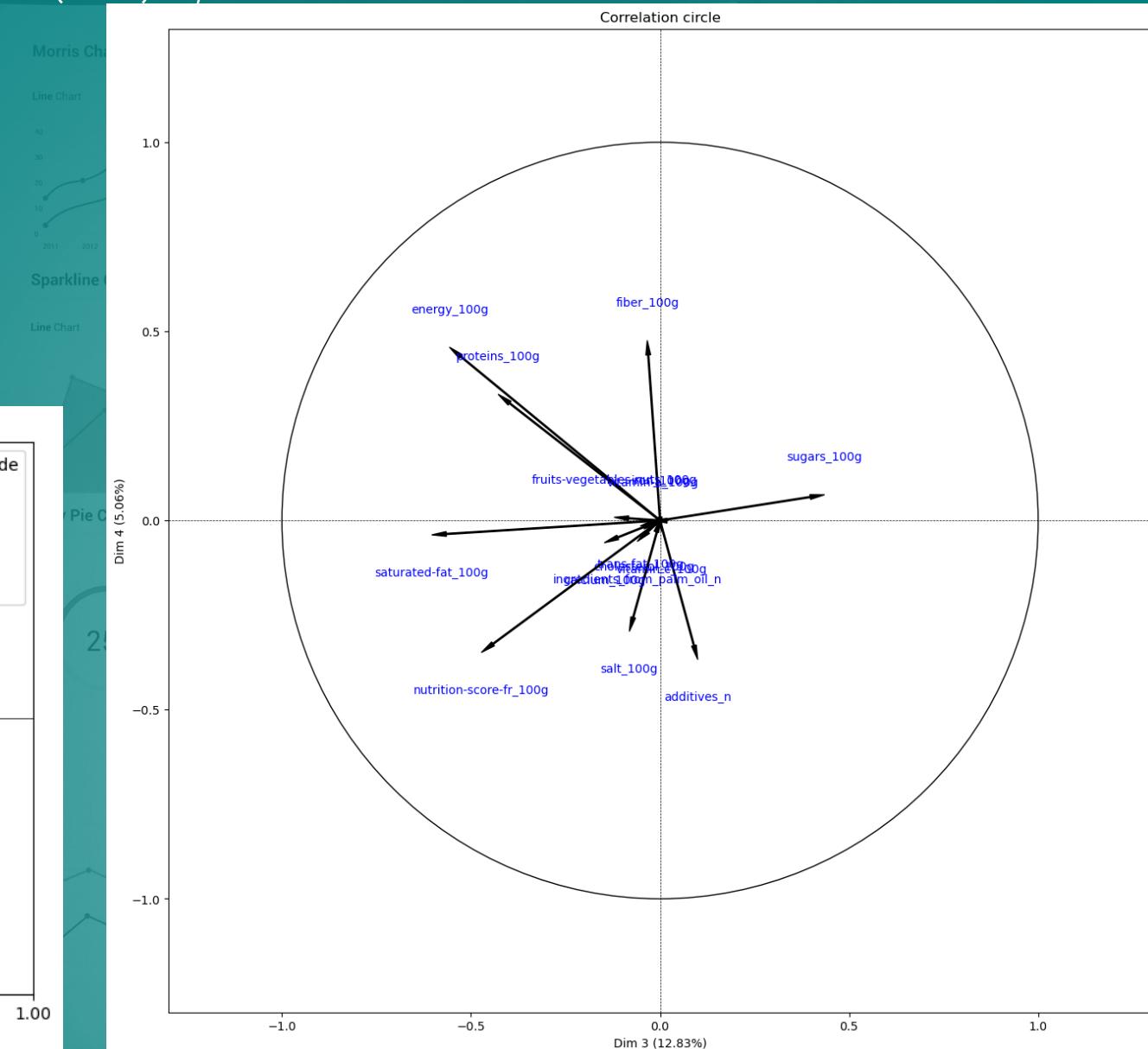
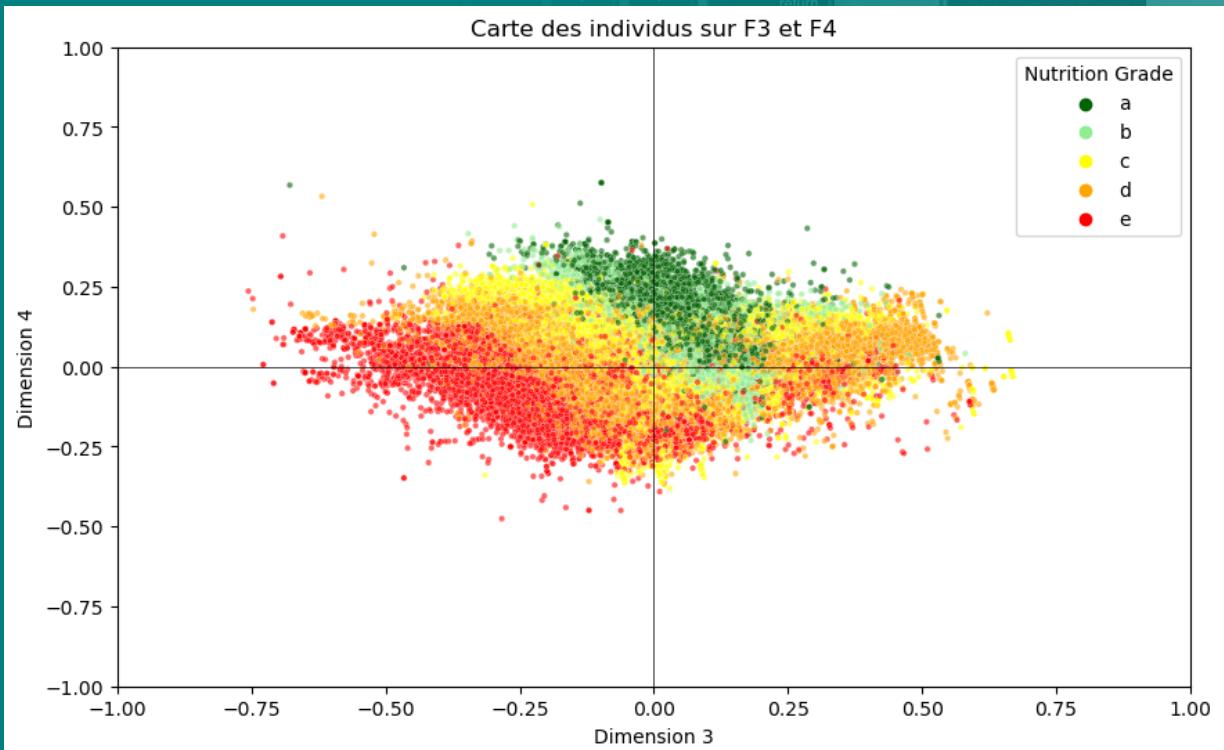


## 5. Analyse multivariée

Analyse en composantes principales (ACP) .2/2

### *Interprétation des composantes F3 et F4*

- F3 : corrélation positive avec les produits sucrés et corrélation négative avec les produits gras, protéinés, énergétique et avec un haut score nutritionnel.
- F4 : corrélation positive avec les produits à forte teneur en fibre et protéines.



## 5. Analyse multivariée

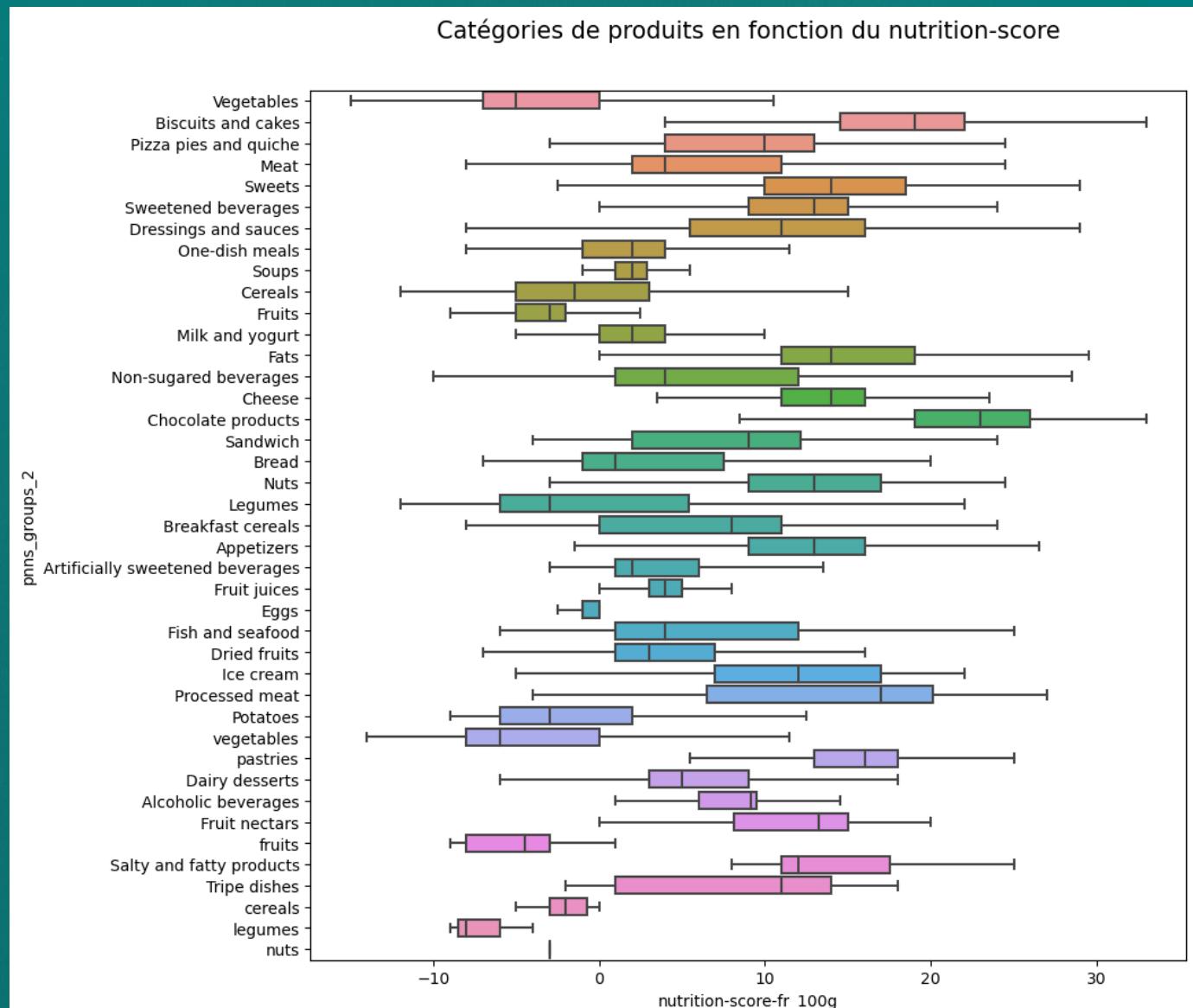
### ANOVA 1/2

#### Objectif de l'ANOVA

Analyser les différences entre les catégories de produits en fonction du nutrition-score.

#### Boxplot

On voit ici la variabilité du nutrition-score par rapport aux catégories de produits.



# 5. Analyse multivariée

## ANOVA . 2/2

- Le modèle explique 53,5% de la variance.
- Les coefficient associés à chaque catégorie nous montre leur impact sur le nutrition-score.
- Le tableau de l'analyse de la variance montre une p-valeur du test de Fisher à 0.
- Les catégories de produits jouent donc un rôle dans la variation du nutrition-score.

Tableau de l'analyse de la variance :

	sum_sq	df	F	PR(>F)
pnns_groups_2	2.434066e+06	40.0	1607.405223	0.0
Residual	2.115074e+06	55870.0	NaN	NaN

OLS Regression Results						
Dep. Variable:	nutrition_score_fr_100g	R-squared:	0.535			
Model:	OLS	Adj. R-squared:	0.535			
Method:	Least Squares	F-statistic:	1607.			
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	0.00			
Time:	19:07:21	Log-Likelihood:	-1.8090e+05			
No. Observations:	55911	AIC:	3.619e+05			
Df Residuals:	55870	BIC:	3.622e+05			
Df Model:	40					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	8.3686	0.497	16.824	0.000	7.394	9.344
pnns_groups_2[T.Appetizers]	4.5279	0.517	8.754	0.000	3.514	5.542
pnns_groups_2[T.Artificially sweetened beverages]	-4.2198	0.630	-6.702	0.000	-5.454	-2.986
pnns_groups_2[T.Biscuits and cakes]	9.6960	0.587	19.132	0.000	8.783	10.689
pnns_groups_2[T.Bread]	-4.8447	0.521	-9.298	0.000	-5.866	-3.823
pnns_groups_2[T.Breakfast cereals]	-2.1100	0.526	-4.014	0.000	-3.140	-1.080
pnns_groups_2[T.Cereals]	-8.1259	0.508	-16.007	0.000	-9.121	-7.131
pnns_groups_2[T.Cheese]	4.4878	0.508	8.832	0.000	3.492	5.484
pnns_groups_2[T.Chocolate products]	12.9413	0.512	25.293	0.000	11.938	13.944
pnns_groups_2[T.Dairy desserts]	-1.8526	0.547	-3.386	0.001	-2.925	-0.780
pnns_groups_2[T.Dressings and sauces]	2.3266	0.511	4.554	0.000	1.325	3.328
pnns_groups_2[T.Dried fruits]	-4.5074	0.583	-7.725	0.000	-5.651	-3.364
pnns_groups_2[T.Eggs]	-8.0808	0.673	-12.004	0.000	-9.400	-6.761
pnns_groups_2[T.Fats]	6.8745	0.525	13.101	0.000	5.846	7.983
pnns_groups_2[T.Fish and seafood]	-2.5582	0.516	-4.961	0.000	-3.569	-1.547
pnns_groups_2[T.Fruit juices]	-4.0128	0.519	-7.730	0.000	-5.630	-2.995
pnns_groups_2[T.Fruit nectars]	3.5660	0.598	5.959	0.000	2.393	4.739
pnns_groups_2[T.Fruits]	-9.8829	0.527	-18.754	0.000	-10.916	-8.850
pnns_groups_2[T.Ice cream]	3.8582	0.553	6.973	0.000	2.774	4.943
pnns_groups_2[T.Legumes]	-7.7876	0.546	-14.268	0.000	-8.857	-6.718
pnns_groups_2[T.Meat]	-2.1678	0.530	-4.093	0.000	-3.206	-1.130
pnns_groups_2[T.Milk and yogurt]	-5.1911	0.510	-10.172	0.000	-6.191	-4.191
pnns_groups_2[T.Non-sugared beverages]	-1.7438	0.514	-3.392	0.001	-2.751	-0.736
pnns_groups_2[T.Nuts]	4.0503	0.561	7.224	0.000	2.951	5.149
pnns_groups_2[T.One-dish meals]	-5.5087	0.505	-10.907	0.000	-6.499	-4.519
pnns_groups_2[T.Pizza pies and quiche]	0.3233	0.574	0.563	0.573	-0.882	1.449
pnns_groups_2[T.Potatoes]	-9.5565	0.804	-11.891	0.000	-11.132	-7.981
pnns_groups_2[T.Processed meat]	6.3711	0.512	12.439	0.000	5.367	7.375
pnns_groups_2[T.Salty and fatty products]	5.0524	1.497	3.376	0.001	2.119	7.986
pnns_groups_2[T.Sandwich]	-0.4968	0.554	-0.897	0.370	-1.582	0.589
pnns_groups_2[T.Soups]	-6.2604	0.574	-10.908	0.000	-7.385	-5.136
pnns_groups_2[T.Sweetened beverages]	3.6135	0.517	6.995	0.000	2.601	4.626
pnns_groups_2[T.Sweets]	5.7134	0.508	11.249	0.000	4.718	6.709
pnns_groups_2[T.Tripe dishes]	-0.4401	1.810	-0.436	0.663	-2.420	1.539
pnns_groups_2[T.Vegetables]	-11.0429	0.517	-21.341	0.000	-12.057	-10.029
pnns_groups_2[T.cereals]	-8.7020	1.845	-4.718	0.000	-12.317	-5.087
pnns_groups_2[T.fruits]	-12.5845	1.053	-11.956	0.000	-14.648	-10.522
pnns_groups_2[T.legumes]	-15.3686	3.587	-4.285	0.000	-22.399	-8.338
pnns_groups_2[T.nuts]	-11.3686	6.173	-1.842	0.066	-23.468	0.730
pnns_groups_2[T.pastries]	7.8470	0.584	12.061	0.000	5.902	8.192
pnns_groups_2[T.vegetables]	-12.4046	0.548	-22.655	0.000	-13.478	-11.331
Omnibus:	1435.320	Durbin-Watson:	1.427			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2030.604			
Skew:	0.288	Prob(JB):	0.00			
Kurtosis:	3.734	Cond. No.	243.			

## 6. Conclusion

### Faisabilité du projet

#### *Analyse des Données Nutritionnelles*

Nous avons effectué une analyse approfondie des données nutritionnelles, y compris la gestion des valeurs manquantes, la détection des corrélations et l'analyse en composantes principales (ACP).

#### *Implications de l'Analyse Univariée et Multivariée*

L'analyse des données nous a permis de mieux comprendre les caractéristiques nutritionnelles des produits. Nous avons observé des corrélations entre les variables, ce qui peut orienter la conception d'une application d'auto-complétion.

#### *Analyse des Catégories de Produits*

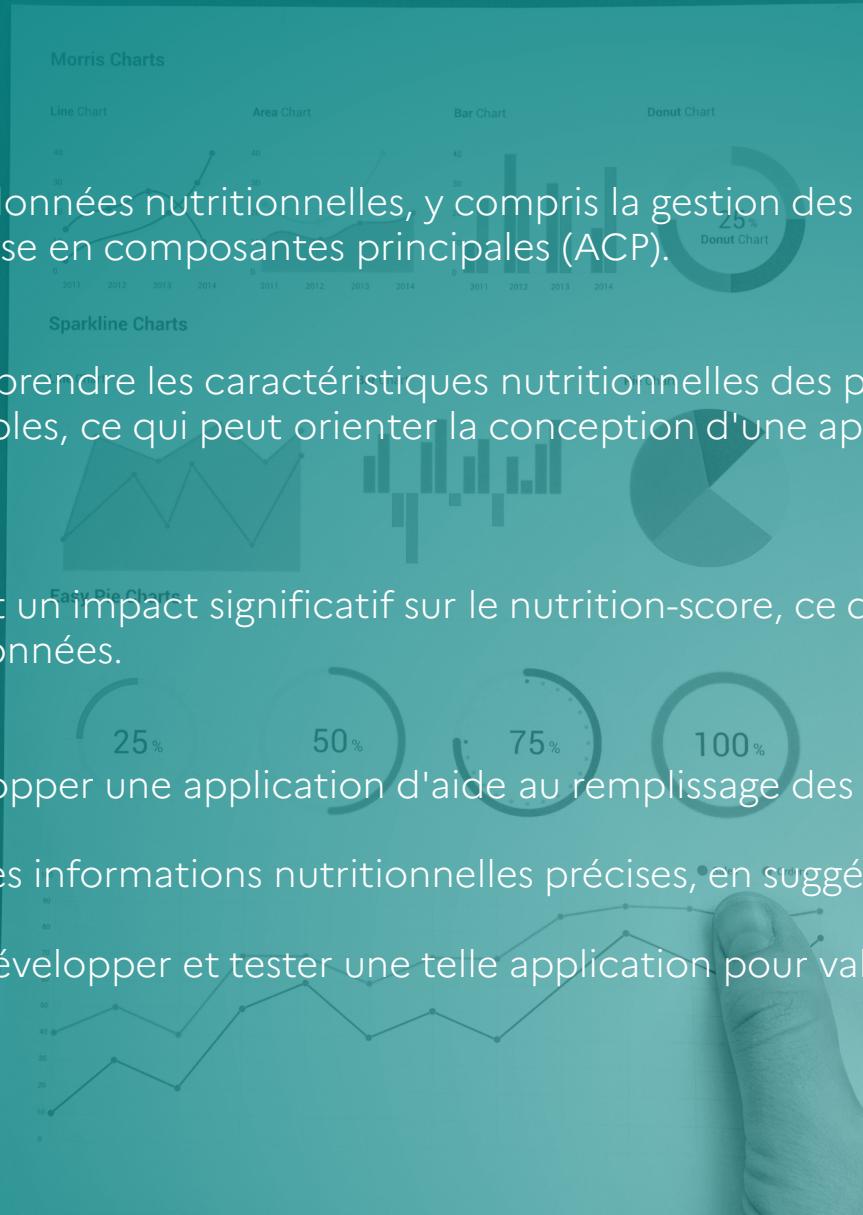
L'ANOVA a révélé que les catégories de produits ont un impact significatif sur le nutrition-score, ce qui renforce l'idée d'une application d'aide au remplissage de données.

#### *Faisabilité de l'Application*

Sur la base de notre analyse, il est faisable de développer une application d'aide au remplissage des données pour les produits alimentaires.

L'application pourrait aider les utilisateurs à saisir des informations nutritionnelles précises, en suggérant des valeurs manquantes à partir de produits similaires.

Les prochaines étapes consisteraient à concevoir, développer et tester une telle application pour valider sa faisabilité et son utilité réelle.



## 6. Conclusion

Respect des 5 grands principes de la RGPD

*Les données présentent dans le jeu de données d'OpenFoodFacts ne présentent pas de risques pour le respect de la RGPD.  
Mais dans le cas d'une conception d'application, il est important de rappeler ces principes :*

### Finalité

Les données collectées servent uniquement à l'analyse nutritionnelle des produits alimentaires.  
Aucune utilisation détournée des données n'est autorisée.



### Proportionnalité et Pertinence

Les données collectées sont proportionnelles et pertinentes au vu de l'objectif.  
Aucune collecte excessive de données n'a lieu.



### Durée de Conservation Limitée

Les données nutritionnelles ne sont conservées que pour la durée de l'analyse.  
Aucune conservation à long terme n'est prévue.



### Informations des Utilisateurs

Les utilisateurs sont informés de la collecte et de l'utilisation de leurs données à des fins d'analyse nutritionnelle.  
Ils ont le droit de donner ou de refuser leur consentement.



### Sécurité et Confidentialité

Les données sont stockées de manière sécurisée pour empêcher tout accès non autorisé.  
La confidentialité des données est garantie.

