



Préparez des données pour un organisme de santé publique





Sommaire

1- Objectif et Mission

2- Idée d'application

3- Nettoyage des données

4- Analyse Univariée

5- Analyse Multivariée

6- Faisabilité de l'application et le respect des principes du RGPD

7- Conclusion



1- Objectif et Mission

Information relative à l'organisation : Santé publique France est un établissement public administratif sous la tutelle du ministère de la Santé. Crée en mai 2016 par ordonnance et décret, A pour mission principale d'améliorer et de protéger la santé des populations.

Objectif : L'agence Santé publique France souhaite enrichir sa **base de données Open Food Facts** en faisant appel à notre entreprise. Cette base de données en accès libre est destinée aux particuliers et aux organisations pour les aider à évaluer la qualité nutritionnelle des produits.

Mission : réalisation d'un projet interne de nettoyage et d'exploration des données pour évaluer la faisabilité d'une application liée à l'alimentation pour **Santé publique France**.

Données :

- Extraits du site Open Food Facts : <https://world.openfoodfacts.org/>
- Guide d'explication des variables disponible à : <https://world.openfoodfacts.org/data/data-fields.txt>



2- Idée d'application

Objectif :

- Créer un système de suggestion ou d'auto-complétion pour faciliter l'ajout de produits à la base de données Open Food Facts.
- Améliorer l'efficacité et la précision de la saisie des informations nutritionnelles des produits.

Fonctionnalités :

- **Suggestion Automatique** : Lorsque les utilisateurs commencent à saisir les informations d'un produit, le système propose des suggestions basées sur les données existantes dans la base.
- **Correction Automatique** : Le système détecte et corrige les erreurs de saisie potentielles, comme les fautes de frappe ou les valeurs incohérentes.
- **Complétion des Champs** : Il complète automatiquement les champs les plus courants et réduit le temps de saisie pour les utilisateurs.
- **Vérification de la Cohérence** : Le système alerte les utilisateurs en cas de valeurs nutritionnelles qui paraissent aberrantes ou incohérentes avec les données existantes.

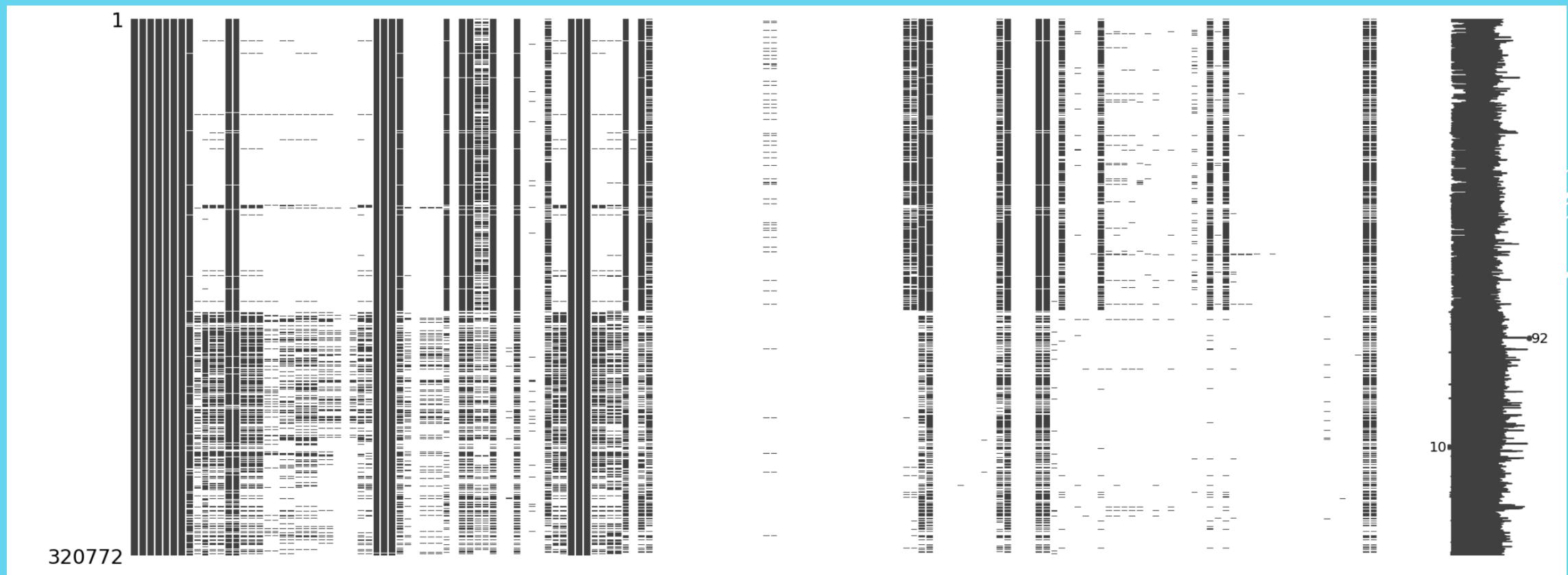




3- Nettoyage des données

Données :

- Un total de **320,772** produits comprenant **162** variables informatives (textes descriptifs et données nutritionnelles).
- Présence de **76%** de données manquantes dans ce jeu de données.





Processus de nettoyage:

Etapes	Action	Nb_Colonne	Nb_Ligne	Pourcentage_NaN
0	Ouverture du Fichier	162	320772	76
1	Suppression des colonnes ayant moins de 50% de leur valeur remplie, sauf 'pnns_groups_1', 'pnns_groups_2'	36	320772	17
2	Suppression des lignes ne contenant pas d'informations sur les ingrédients	36	262768	9
3	Analyse des variables pour les informations générales	31	259380	11
4	Analyse des variables tags	28	259380	12
5	Analyse des variables Igredients	28	259380	12
6	Analyse des variables Données diverses	22	259380	10
7	Analyse des variables apports nutritionnels	21	259206	10
8	Traitement des colonnes textuelles et élimination des produits sans groupe PNNS2	21	259206	4
9	Génération des variables associées à l'EnvironnementScore.	20	259206	3
10	Utilisation de la méthode KNN pour remplir les colonnes quantitatives et suppression des lignes avec 'Nom Communiqué'.	20	192980	0
11	Suppression des valeurs aberrantes, incluant les valeurs négatives, celles supérieures à 100, ainsi que les sommes totales dépassant 100.	20	191256	0
12	Élimination et Suppression manuelle des outliers (valeurs aberrantes).	20	191256	0



Description des étapes du processus de nettoyage:

Étape 1 : Nous avons procédez à l'ouverture du fichier afin d'examiner les variables disponibles, ainsi que pour comprendre le nombre de lignes et de colonnes. Cette étape s'est révélée cruciale pour les analyses ultérieures.

Étape 2 à Étape 8 : Nous avons éliminé les colonnes avec moins de 50% de valeurs remplies, à l'exception des colonnes PNNS Groupe 1 et PNNS Groupe 2, qui sont essentielles pour le Programme National Nutrition Santé. De plus, à l'**Étape 3**, nous avons supprimé les lignes sans informations sur les ingrédients. De l'**Étape 4 à l'Étape 8**, nous avons ensuite procédé à l'analyse des variables : d'abord les informations générales, puis les informations TAG, les informations sur les ingrédients, les données diverses, et enfin, les données sur l'apport nutritionnel.

À partir de l'**Étape 9** jusqu'à l'**Étape 13**, plusieurs processus ont été réalisés pour le nettoyage et la préparation des données :

À l'**Étape 9**, nous avons traité les colonnes textuelles en remplaçant les valeurs qualitatives manquantes par 'Non communiqués' et en éliminant les produits sans groupe PNNS Groupe 2 (Programme National Nutrition Santé). Ensuite, à l'**Étape 10**, nous avons généré les variables associées à l'environnement SCORE.

À l'**Étape 11**, les valeurs NaN dans les colonnes quantitatives ont été remplies en utilisant la méthode KNN. Simultanément, les lignes contenant 'Non communiqués' dans les variables qualitatives ont été supprimées. Puis, à l'**Étape 12**, nous avons éliminé les valeurs aberrantes, y compris les valeurs négatives, les valeurs supérieures à 100, ainsi que les sommes totales dépassant 100.

Enfin, à l'**Étape 13**, les outliers ont été éliminés manuellement en remplaçant certaines valeurs aberrantes par NaN. Ensuite, nous avons rempli les valeurs manquantes en utilisant la médiane des valeurs en dessous du seuil autorisé.



4- Analyse Univariée

Paramètres descriptifs des données:

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g
count	191256.000000	191256.000000	191256.000000	191256.000000	191256.000000	191256.000000	191256.000000	191256.000000	191256.000000
mean	1180.046926	13.300748	4.862029	33.284227	14.993143	2.759077	7.806331	1.242137	0.489030
std	752.511728	15.936571	7.261502	28.214083	19.652709	4.208184	7.930802	3.825096	1.505938
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	458.000000	0.880000	0.000000	7.080000	1.400000	0.000000	2.040000	0.111760	0.044000
50%	1197.000000	7.140000	1.790000	24.290000	5.100000	1.448000	5.710000	0.683260	0.269000
75%	1715.000000	21.430000	7.140000	59.000000	23.200000	3.600000	10.870000	1.400000	0.551181
max	4657.000000	100.000000	100.000000	100.000000	100.000000	100.000000	86.000000	100.000000	39.370079

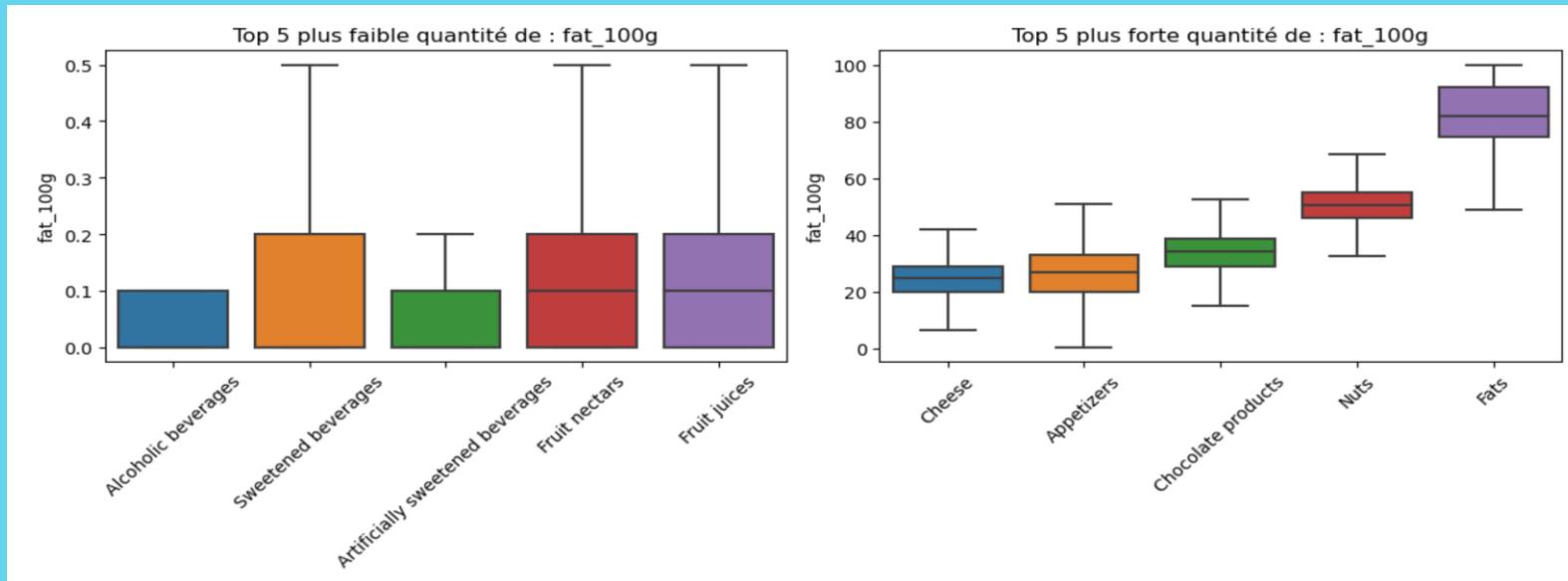
- Il y a des zéros (0), ce qui est normal car tous les produits ne contiennent pas forcément tous les ingrédients. Cela peut indiquer l'absence d'un ingrédient dans un produit.



Analyse par catégorie :

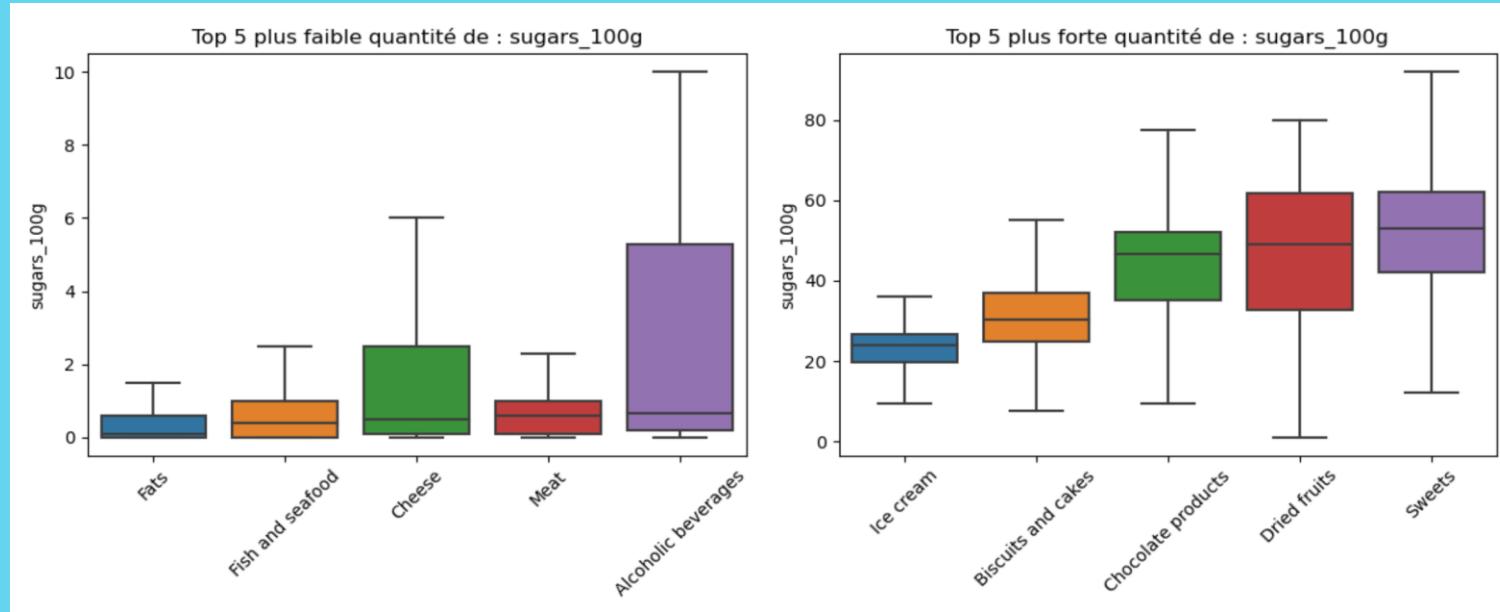
	saturated-fat_100g	energy_100g	sodium_100g	salt_100g	proteins_100g	sugars_100g	fat_100g	carbohydrates_100g	fiber_100g
Maximum	Fruit nectars	Fruit nectars	Fruit nectars	Fruit nectars	Fruit nectars	Fruit nectars	Eggs	Fruit nectars	Tripe dishes
Minimum	Salty and fatty products	Salty and fatty products	Tripe dishes	Tripe dishes	Tripe dishes	Ice cream	Alcoholic beverages	Potatoes	Alcoholic beverages
Moyenne Maximum	Fats	Fats	Processed meat	Processed meat	Processed meat	Sweets	Fats	Breakfast cereals	Salty and fatty products
Moyenne Minimum	Fats	Fats	Processed meat	Processed meat	Processed meat	Sweets	Fats	Breakfast cereals	Salty and fatty products

- La ligne minimale n'est pas très précise car plusieurs produits de différentes catégories peuvent être à zéro, tandis que les moyennes offrent une perspective plus intéressante et représentative.



Observations :

- Les aliments les plus riches en matières grasses sont les graisses, les noisettes et le chocolat.
- Les aliments les moins riches en matières grasses sont les jus de fruits et les alcools.

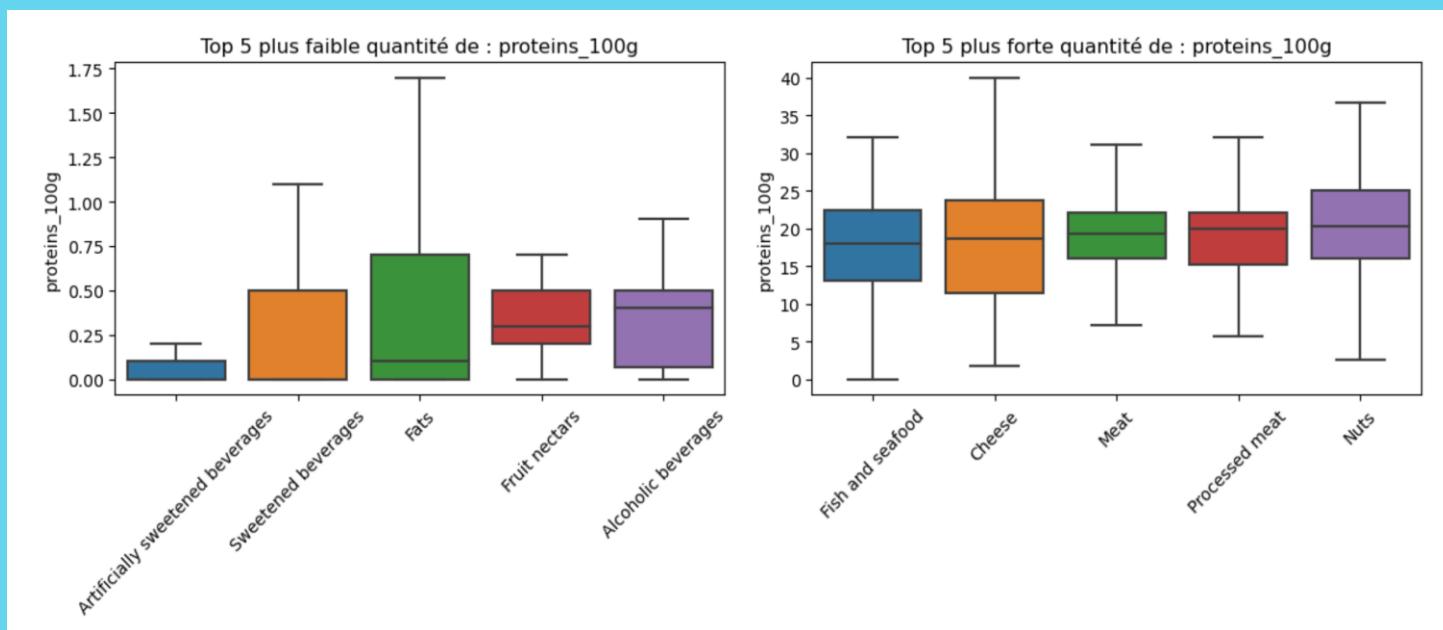
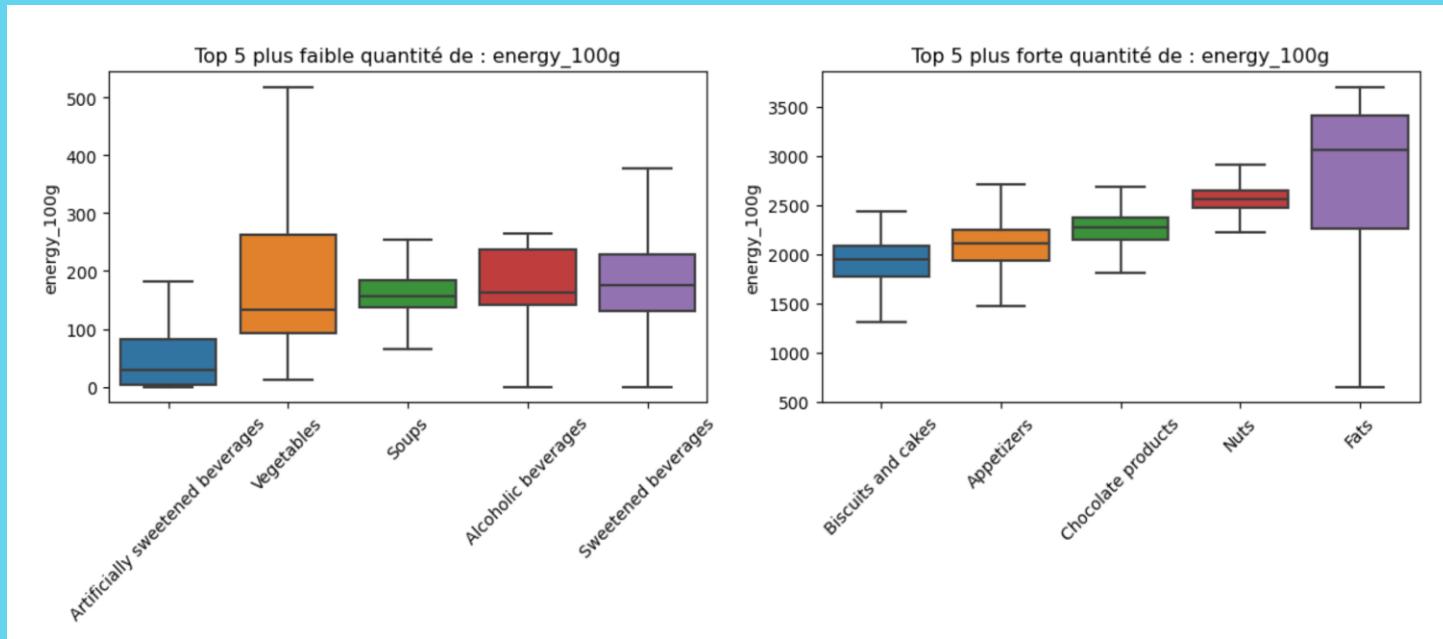


Observations :

- Les aliments les plus sucrés comprennent les bonbons, les produits à base de chocolat et les fruits secs.
- Les aliments les moins sucrés incluent les graisses, les poissons et les boissons alcoolisées.

Observations :

- Les aliments les plus riches en sel comprennent les viandes transformées et les produits salés et gras.
- Les aliments les moins riches en sel sont les boissons à base de fruits ou alcoolisées.



Observations :

- Les aliments les plus riches en énergie sont les graisses, les noisettes et le chocolat.
- Les aliments les moins riches en énergie sont les légumes et les boissons sucrées artificielles.

Observations :

- Les aliments les plus riches en protéines sont les noisettes et les viandes transformées.
- Ceux qui contiennent le moins de protéines sont les graisses, les boissons sucrées artificielles et alcoolisées.



5- Analyse Multivariée

- Analyse de la corrélation entre les variables nutritionnelles et le Nutrigrade.

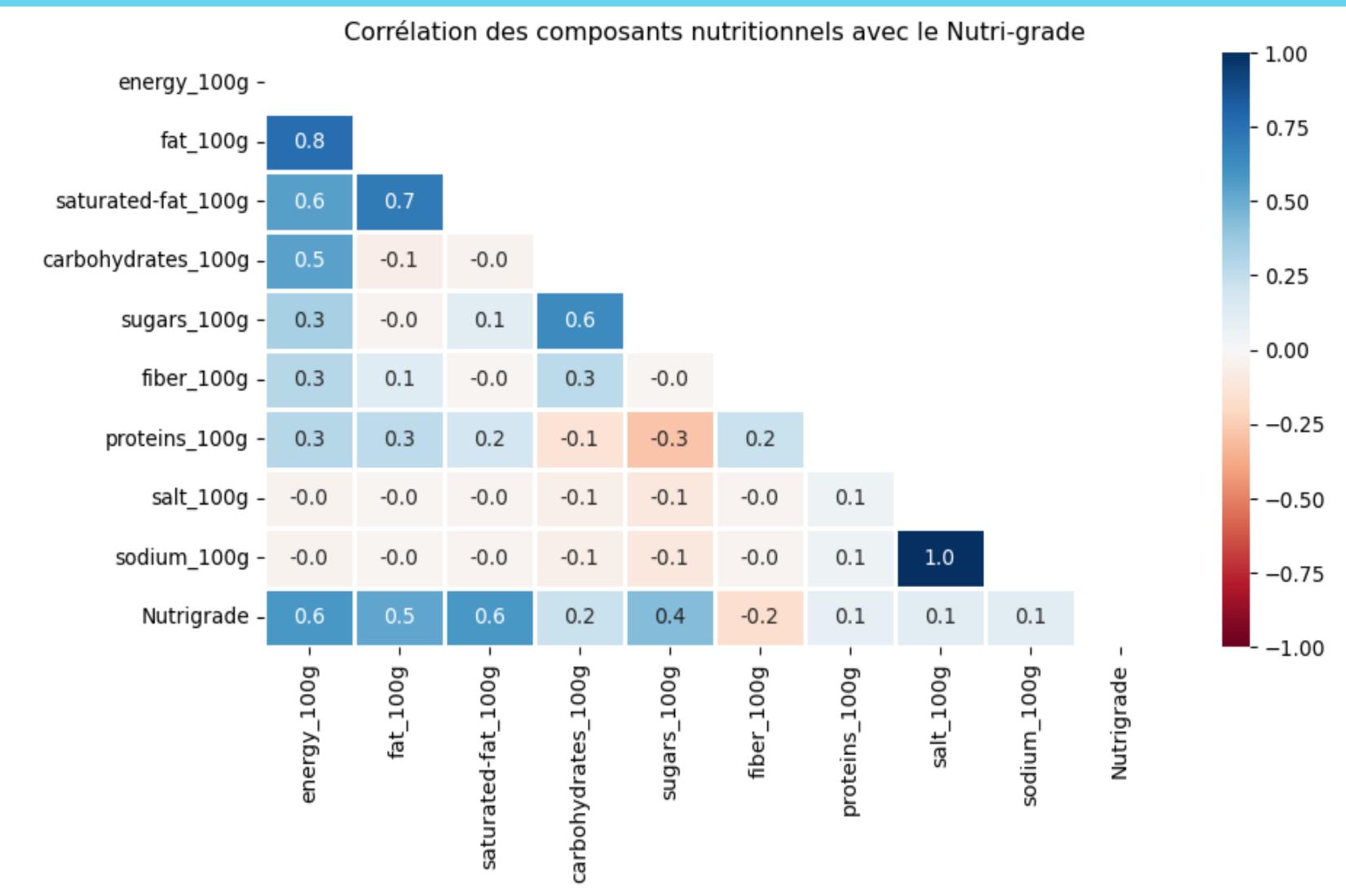
```
Nutrigrade          1.000000
saturated-fat_100g  0.581403
energy_100g         0.578495
fat_100g            0.529151
sugars_100g         0.434140
carbohydrates_100g 0.226301
sodium_100g         0.113480
salt_100g           0.113480
proteins_100g        0.091042
fiber_100g          -0.151530
Name: Nutrigrade, dtype: float64
```

Observations :

- Le Nutrigrade est principalement corrélé avec les graisses et le sucre.
- Il est également corrélé avec l'énergie, mais cette variable est calculée principalement à partir des valeurs de graisses.
- En revanche, bien que le calcul du Nutri-score prenne en compte les fibres et les protéines, il semble ne pas y avoir de corrélation avec le Nutri-score.
- Pour résumé, les colonnes fibres, protéines et sel ne présentent pas de corrélation apparente avec le Nutrigrade.



- Analyse de la corrélation entre toutes les variables :

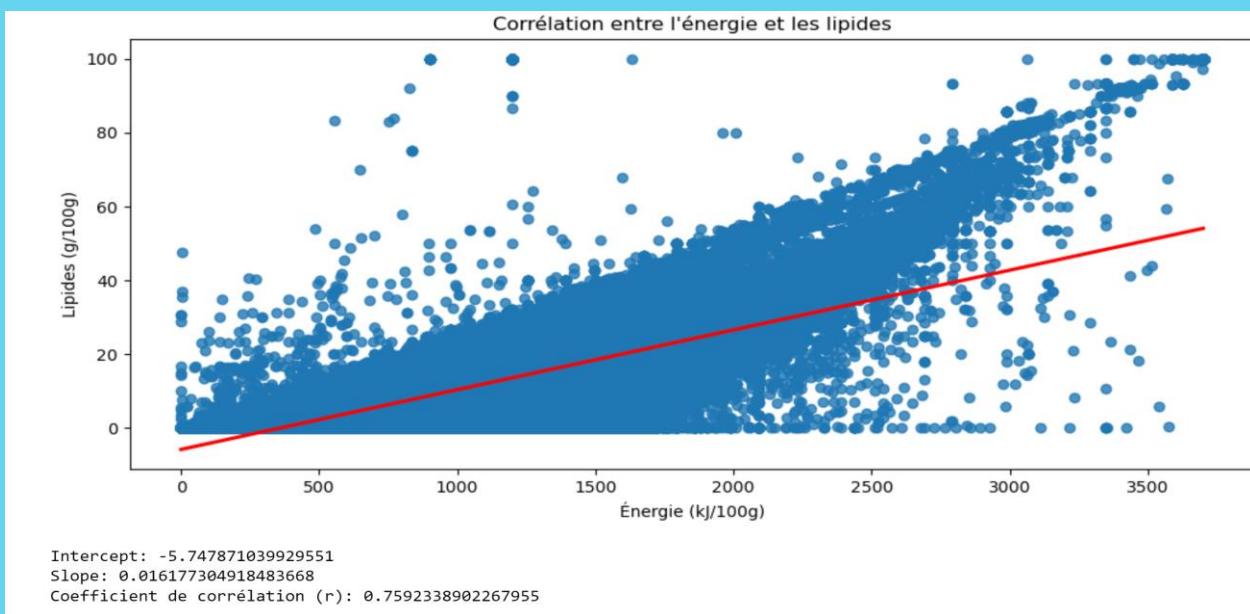
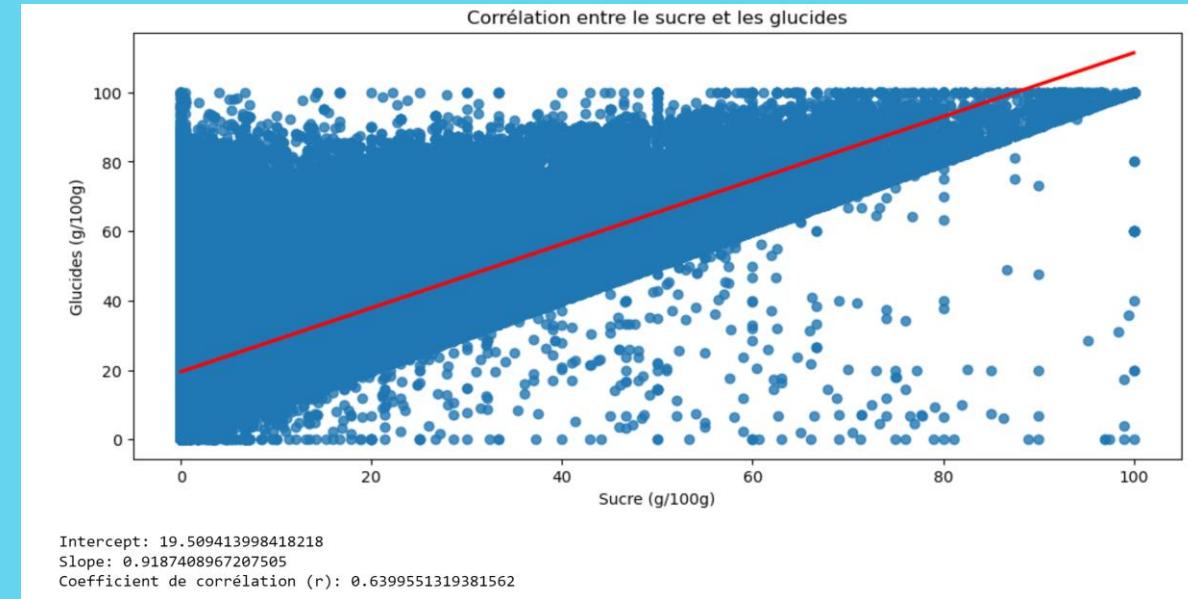
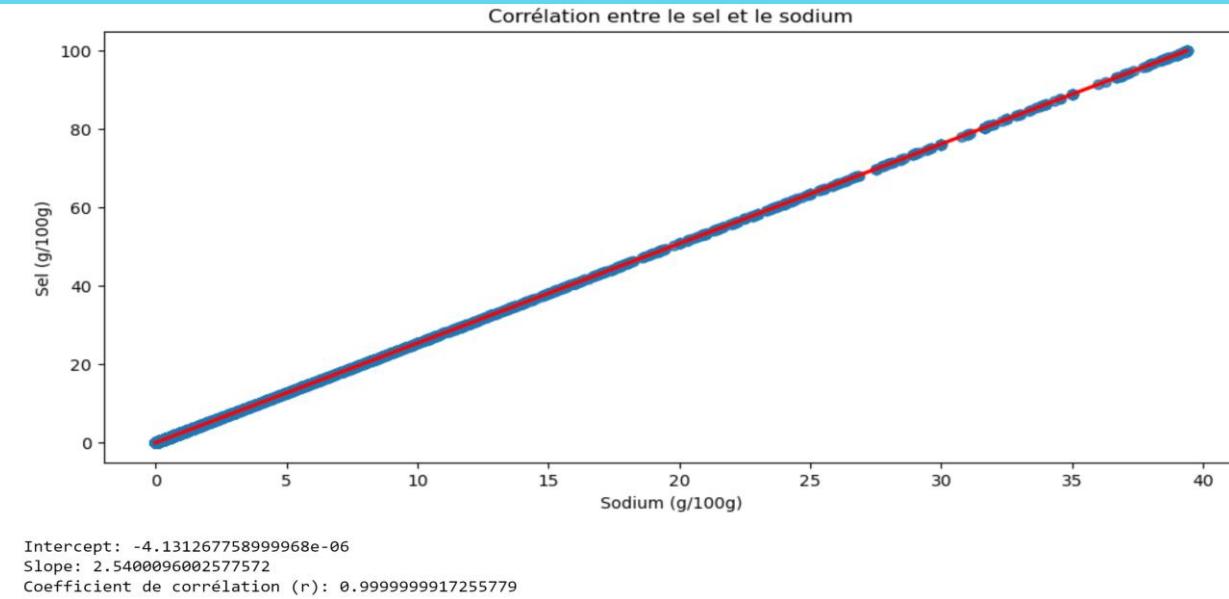


Observations :

- Outre les corrélations avec le Nutrigrade.
- Une forte corrélation est observée entre le gras et le gras saturée (**0.7**).
- Une corrélation parfaite est présente entre le sel et le sodium (**1**).
- Une corrélation est observée entre le gras et l'énergie (**0.8**).
- De même, une corrélation est constatée entre le sucre et les glucides (**0.6**).



▪ Analyse visuelle de ces variables :

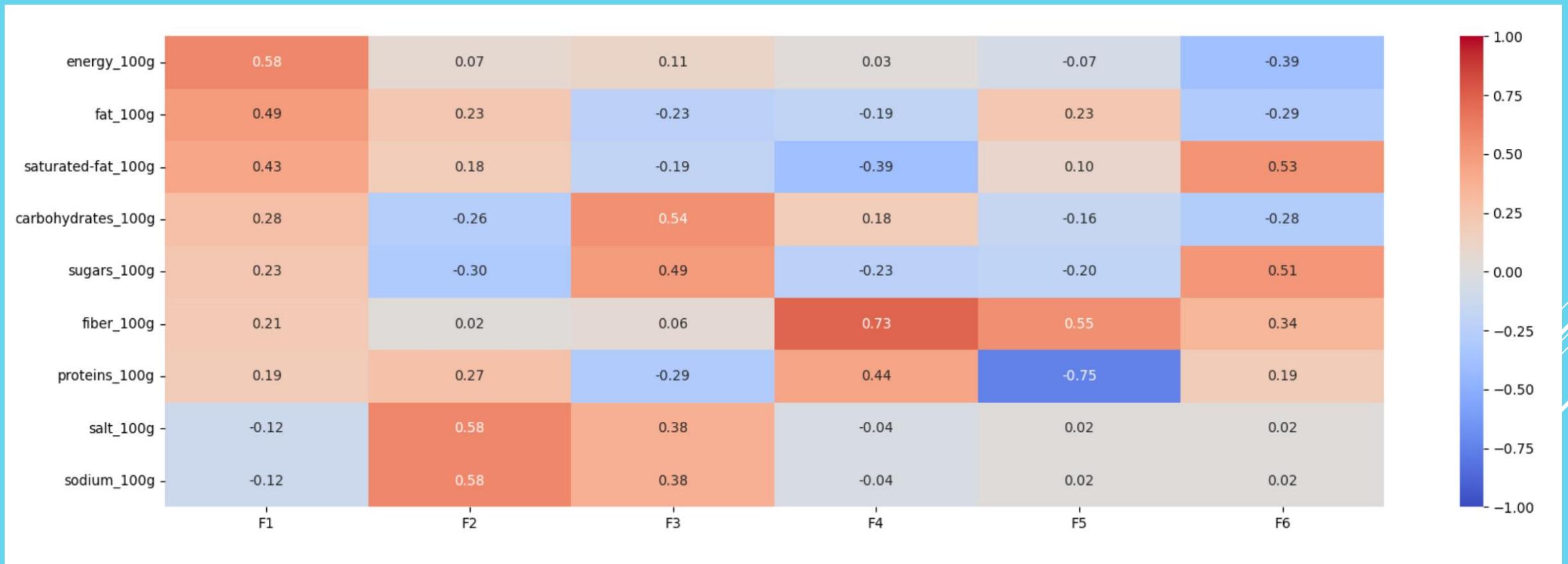


Observations :

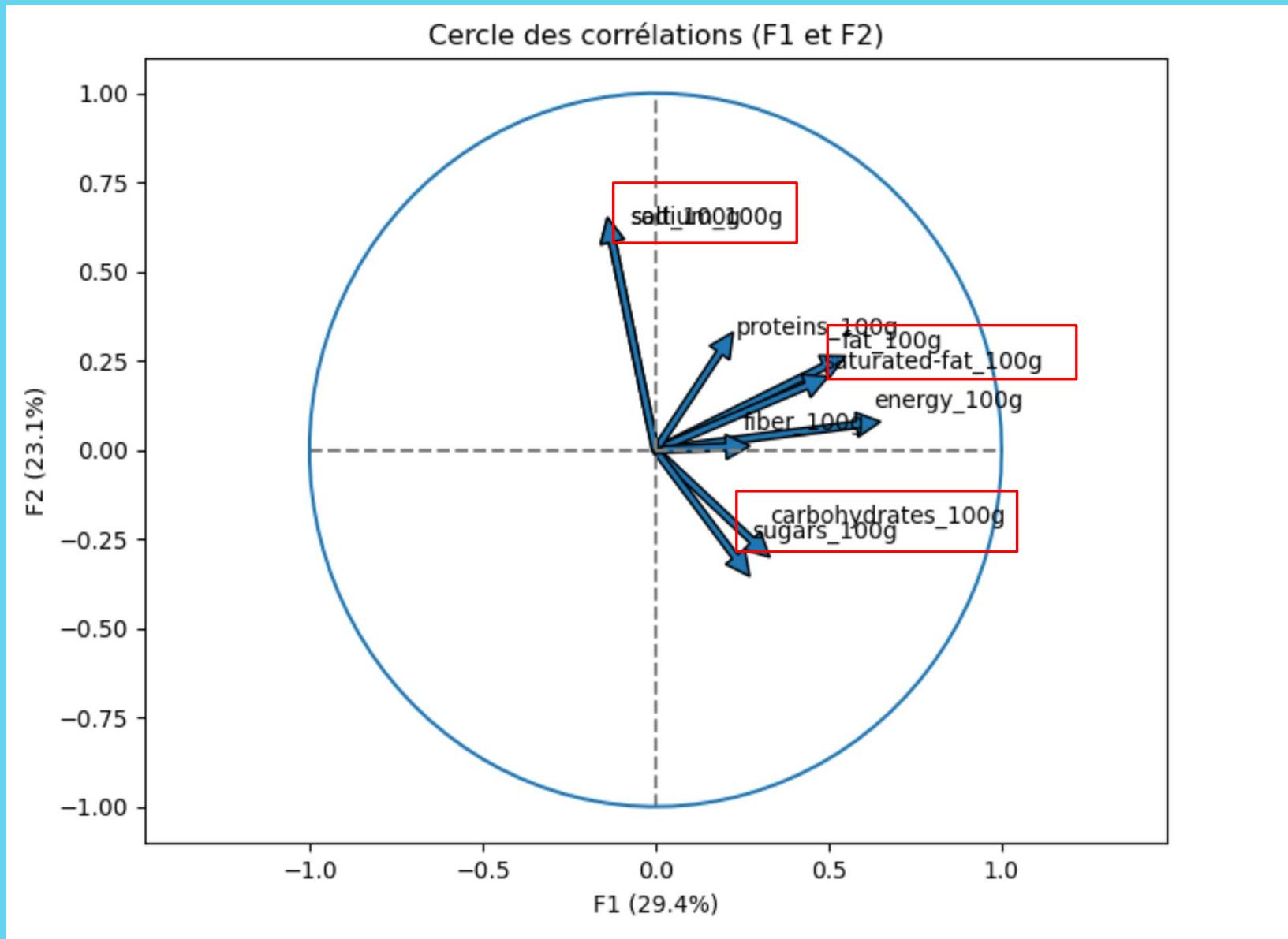
- Ces graphiques mettent en évidence une corrélation très forte entre :
- Le sel et le sodium
- Les graisses et l'énergie
- Le sucre et les glucides



- Analyse de la corrélation des variables à l'aide des composantes principales ou des cercles de corrélations :



- Pour clarifier ce tableau, nous allons nous référer aux cercles de corrélations ci-dessous afin de mieux comprendre les corrélations avec les différentes composantes.

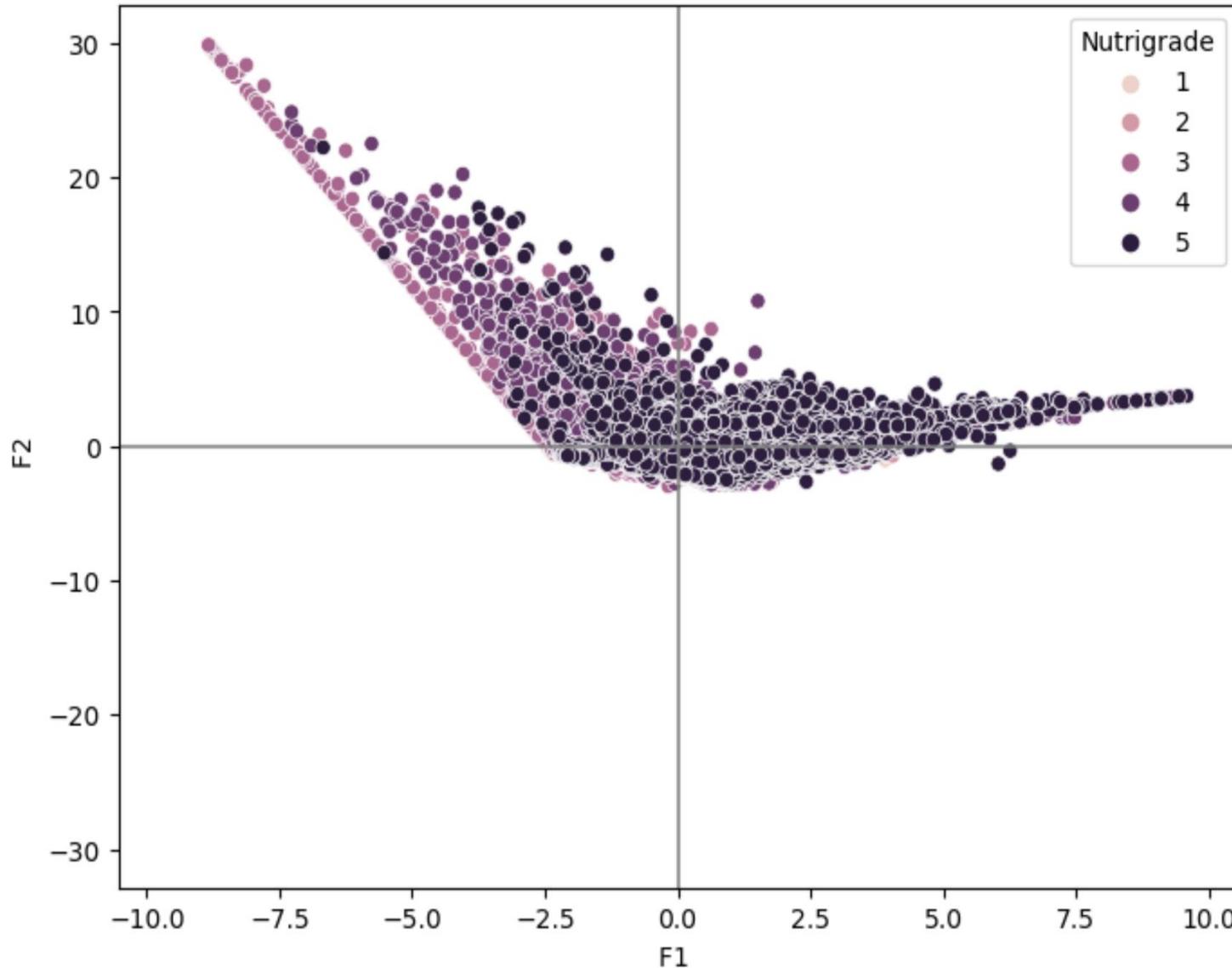


Observations :

- Confirmation des diverses corrélations entre les variables.
- Corrélation inverse entre le sucre et le sel.
- Dans ce graphique, on peut voir les variables "Salt" et "Sodium" sont fortement corrélées négativement avec la première composante principale (PC1), tandis que les variables "gras", "gras saturé" et "Energie" sont corrélées positivement avec la deuxième composante principale (PC2)
- Les variables "protéine", "fibre", "glucide" et "sucre" ont une corrélation plus faible avec les composantes principales, car leurs flèches sont plus courtes et plus proches du centre du cercle.



Projection des ingrédients (sur F1 et F2)



Observations :

- Nous pouvons voir l'énergie corrélé positivement à la composante F2.
- De plus, les couleurs foncées indiquent une augmentation de l'énergie, ce qui signifie que les couleurs foncées ont une valeur énergétique plus élevée que les autres composants nutritionnels.



□ Test statistique ANOVA (analyse de variance)

- `pnns_groups_2`
- `proteins_100g`

Résultats du test ANOVA:

Statistique F : 781.3463322719117

P-valeur : 0.0

- Le test **ANOVA** révèle une différence significative entre les moyennes des groupes '`proteins_100g`' pour chaque catégorie de '`pnns_groups_2`' ($F = 781.34$, $p < 0.001$), indiquant que le type de '`pnns_groups_2`' a un impact notable sur les valeurs de '`proteins_100g`'.
- La statistique F élevée ($F = 781.34$) et la p-valeur très faible ($p < 0.001$) confirment qu'il existe des variations significatives des '`proteins_100g`' entre les différentes catégories de '`pnns_groups_2`'.



6- Faisabilité de l'application

Avantage :

- **Gain de Temps** : Les utilisateurs peuvent ajouter des produits plus rapidement et facilement.
- **Réduction des Erreurs** : Diminue les risques d'erreurs de saisie et de données manquantes.
- **Amélioration de la Base de Données** : Une base de données plus complète et précise, bénéfique pour les particuliers et les organisations utilisant Open Food Facts.
- **Facilitation de la Contribution** : Encourage davantage d'utilisateurs à contribuer à la base de données grâce à une expérience utilisateur améliorée.

Approche :

- **Nettoyage et Exploration des Données** : En premier lieu, l'équipe se concentre sur le nettoyage des données existantes dans la base Open Food Facts pour assurer leur qualité.
- **Modélisation du Système** : Développer un algorithme de suggestion et d'auto-complétion basé sur les données nettoyées.
- **Intégration à Open Food Facts** : Intégrer le système de suggestion dans l'interface d'ajout de produits sur Open Food Facts.

Faisabilité:

- L'idée est réalisable en se basant sur les données disponibles sur Open Food Facts.
- Une équipe technique peut développer cet outil en utilisant des technologies de traitement du langage naturel (NLP) et de machine learning pour les suggestions et les corrections automatiques.
- En résumé, cette idée d'application de système de suggestion pour Open Food Facts répond aux besoins exprimés par l'agence Santé publique France. Elle vise à simplifier et améliorer le processus d'ajout de produits à la base de données, tout en garantissant une meilleure qualité et une précision accrue des informations nutritionnelles.



Respect des principes du RGPD (Règlement général sur la protection des données)



Points à retenir :

Le projet respecte les principes clés du RGPD en assurant la transparence, la minimisation des données, l'exactitude, la limitation des finalités et de la conservation. Ces pratiques garantissent un traitement responsable des données nutritionnelles tout en préservant la confidentialité et la sécurité des utilisateurs. La publication sur le site Open Food Facts contribue à la transparence du processus et démontre l'engagement envers le respect des normes de protection des données.

7- Conclusion

Analyse critique :

- Étant alimentée par le public, la base de données présente des erreurs d'informations fréquentes.
- Le nombre de produits répondant à nos critères est limité, malgré un nettoyage approfondi.
- Il y a des possibilités d'amélioration pour les imputations de données manquantes.

Points à explorer :

- Une vérification préalable et une organisation des données seraient préférables.