PROJET 3

Concevez une application au service de la santé publique

PRESENTATION

1- Présentation de l’ idée d’application.

2- Opérations de nettoyage effectuées.

3- Description et analyse univariée des différentes variables importantes avec les visualisations associées.

4- Analyse multivariée et résultats statistiques associés, en lien avec votre idée d’application.

5- Conclusion

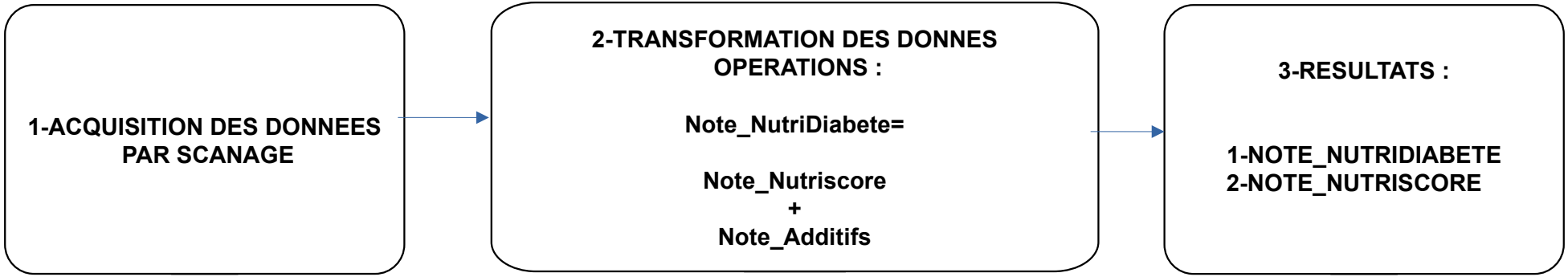
1-Appel à Projet et Présentation de l’ idée d’application.

Appel à Projet

L'agence "Santé publique France" a lancé un appel à projet
pour trouver des idées innovantes d’applications en lien avec l'alimentation.

Idée d’ Application

Nous assistons, ces dernières années à un nombre croissant de cas de **diabète**.
L’idée est de suggérer une application permettant aux personnes diabétiques de choisir des produits compatibles à leur état en évitant entre autre ceux à forte charge glycémique et ou contenant certains additifs spécifiques pouvant aggraver leur état. Pour cela nous nous basons sur le Nutriscore, sytème existant de sélection de produits.


1-ACQUISITION DES DONNEES PAR SCANAGE
2-TRANSFORMATION DES DONNES OPERATIONS :
Note_NutriDiabete=
Note_Nutriscore
+
Note_Additifs
3-RESULTATS :
1-NOTE_NUTRIDIABETE
2-NOTE_NUTRISCORE

- Le Schéma ci-dessus donne une représentation architecturale de l'application proposée :

- **Calcul Note NutriDiabète**

- **Note_NutriDiabete= Note_Nutriscore + Note_Additifs**

- **NutriDiabete**=note à calculer.

- **Note_NutriScore**=note fournie dans le dataset.

- **Note_Additifs** est fonction :

- -du nombre d'additifs présents dans le produit.

- -de la présence d' additifs spéciques reconnus comme nocifs pour les diabétiques (pour ceux-là la note est maximale et fixée à 99.

- **NB** : Plus la note est élevée plus le produit est
- Nocif.

2- Opérations de nettoyage effectuées.

- **A-Jeu de donnée**
  - SOURCES:https://world.openfoodfacts.org/data
  - Dimension
    - Dimension Nouveau Jeu De Donnée: **(320772, 21)**
    - Dimension Ancien Jeu De Donnée: **(320772, 162)**
    - **Nous avons retenons 21 variables sur 162.**

- **B- Il y a 5 types de variables**

  - 1-Informations Générales
  - 2-Tags
  - 3-ingrédients
  - 4-Misc. Data
  - 5-nutrition facts:(Apports nutritionnels)
  -

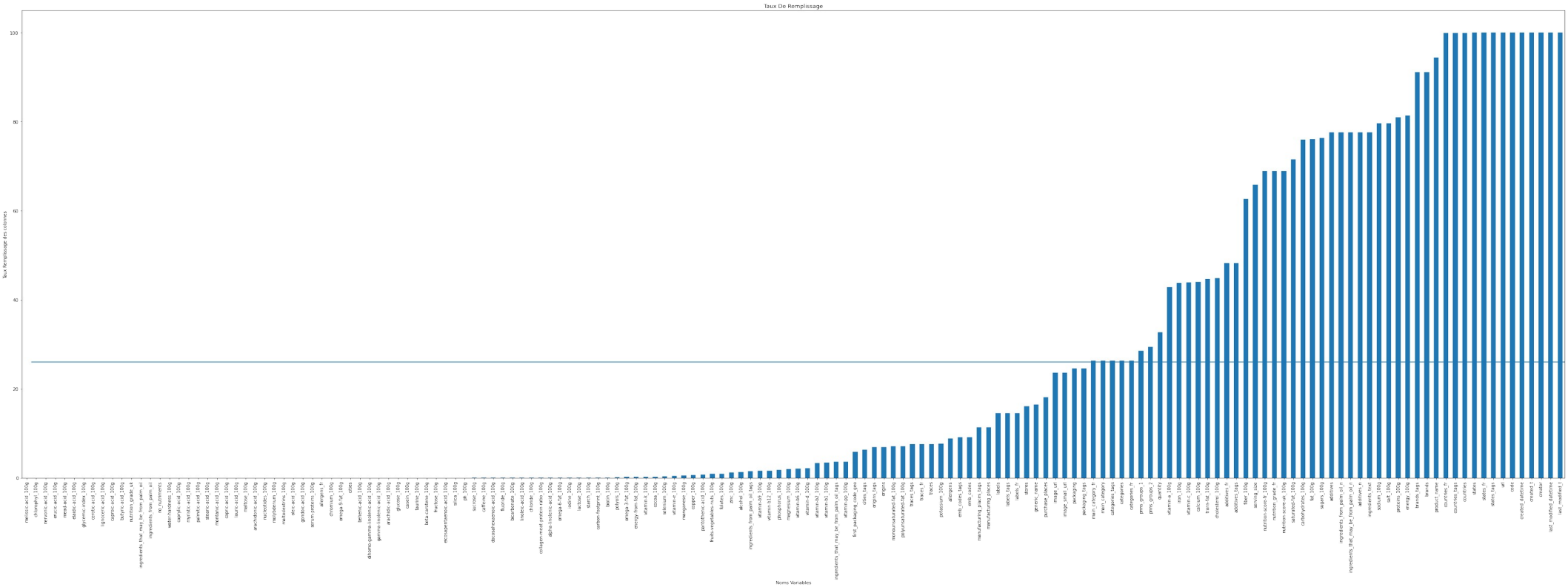
-

C- Selection de variables pertinentes

Retenues :Variables remplies à plus de 26 % :

Index(['product_name', 'categories_fr', 'countries_fr', 'additives_fr', 'nutrition_grade_fr', 'pnns_groups_1', 'energy_100g', 'fat_100g', 'saturated-fat_100g', 'trans-fat_100g', 'carbohydrates_100g', 'fiber_100g', 'proteins_100g', 'salt_100g', 'nutrition-score-fr_100g'**],** **dtype='object')**

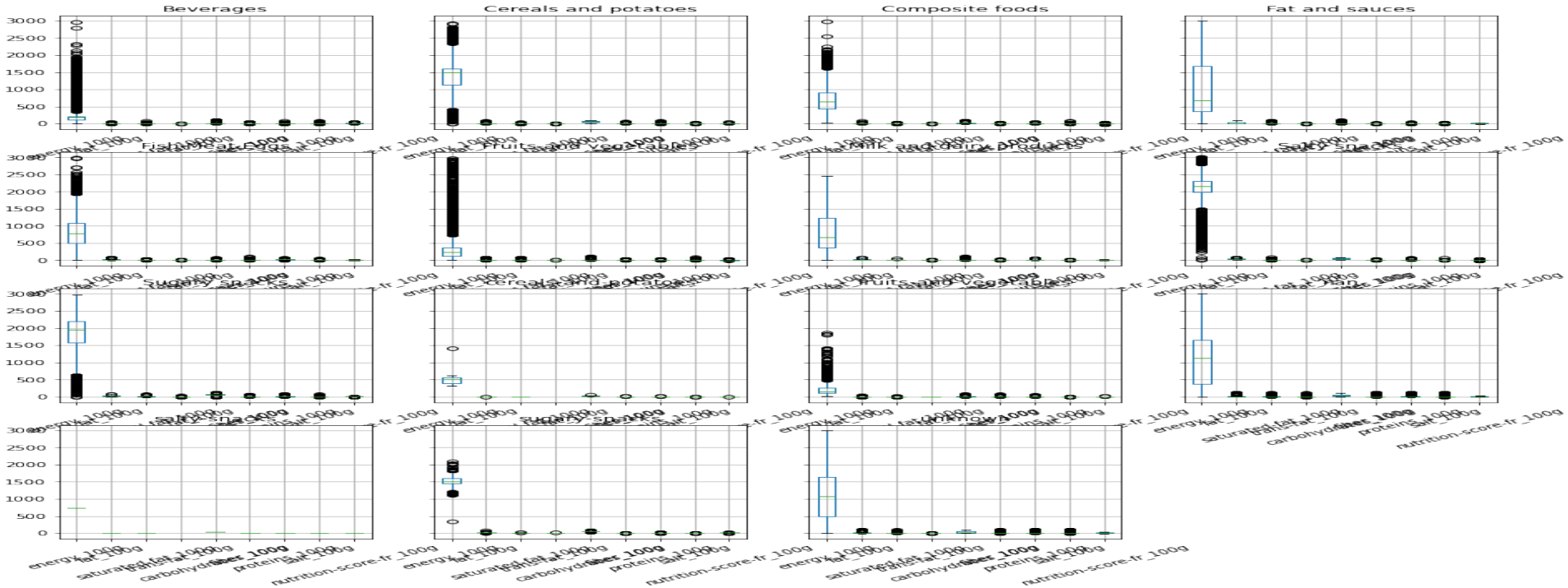
Graphique taux de remplissage :
droite horizontale limite fixée à y=26 % taux de remplissage



- **D-Suppression des doublons**
  - Très peu de doublons trouvés.

- **E-Traitement des outliers.**

  ~ A été appliquée à chaque groupe, la méthode interquartile.



- **F-Valeurs abberrantes**
  - Valeurs des nutriments vérifiées pour qu elles soit comprise entre 0 et 100 Pour les variables
  - fat_100g
  - saturated-fat_100g
  - trans-fat_100g
  - carbohydrates_100g
  - fiber_100g
  - proteins_100g
  - salt_100g

  - Les limites de la variable energie_100g etant 0 et 3000

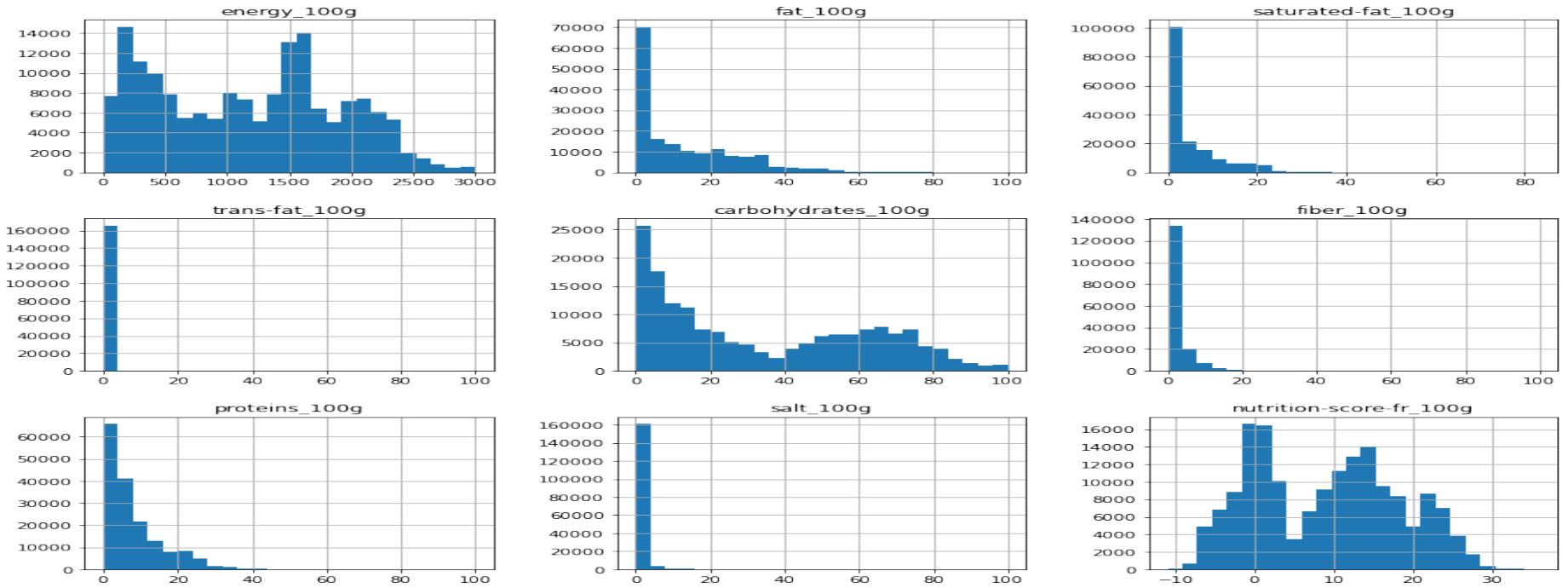
- **G-Valeurs manquantes**
- Remplacement des valeurs manquantes par: 0, moy ou mediane

- **H-Variables et taille du fichier final**

```
Index(['product_name', 'categories_fr', 'countries_fr',
       'additives_fr','nutrition_grade_fr', 'pnns_groups_1',
       'energy_100g', 'fat_100g','saturated-fat_100g', 'trans-
       fat_100g', 'carbohydrates_100g','fiber_100g',
       'proteins_100g', 'salt_100g', 'nutrition-score-fr_100g'],
      dtype='object')
```

3- Description et analyse univariée des différentes variables importantes avec les visualisations associées.

A-Histogramme



REMARQUE

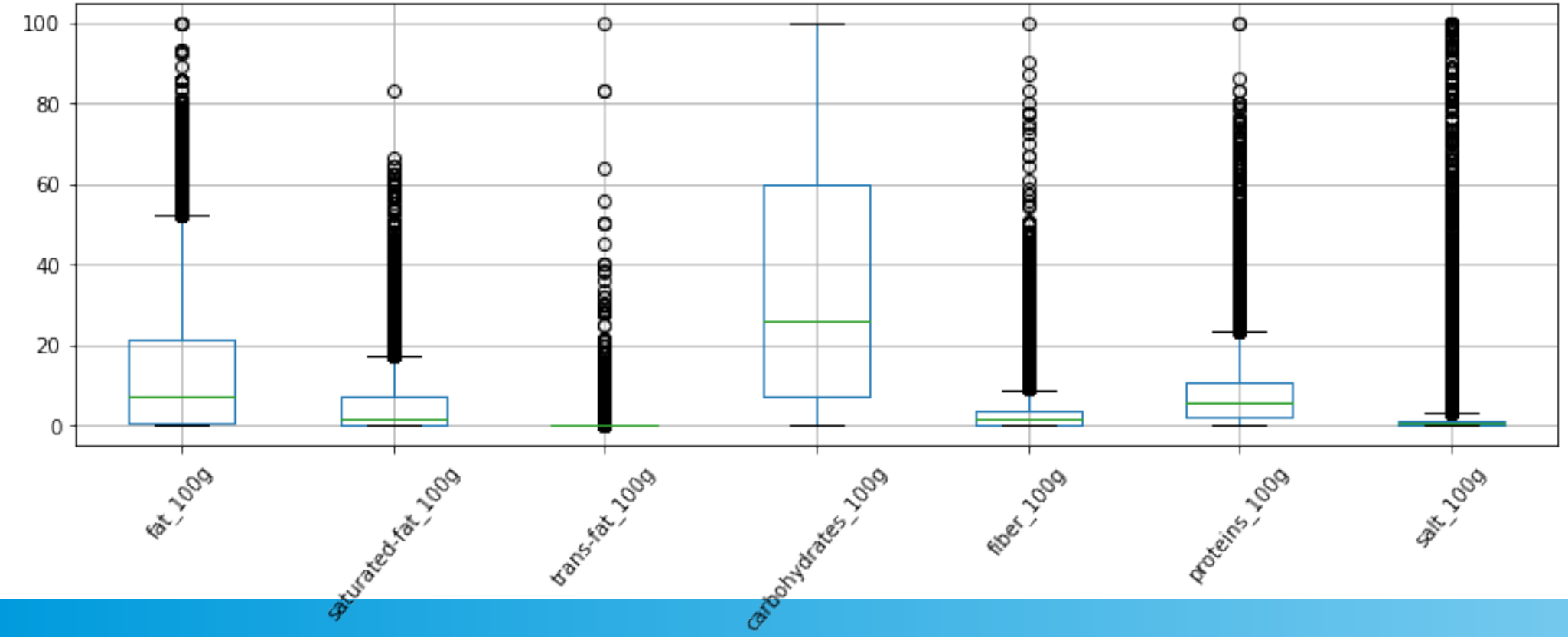
- 1/ Distributions bimodales pour les distributions des variables

  - energy_100g,
  - carbohydrate,
  - nutrition-score-fr_100g

- 2/ Skewness à droite pour
  - fat_100g,
  - saturated-fat_100g
  - fiber_100
  - protein_100

- 3/ Kurtosis important pour l ensemble des variables

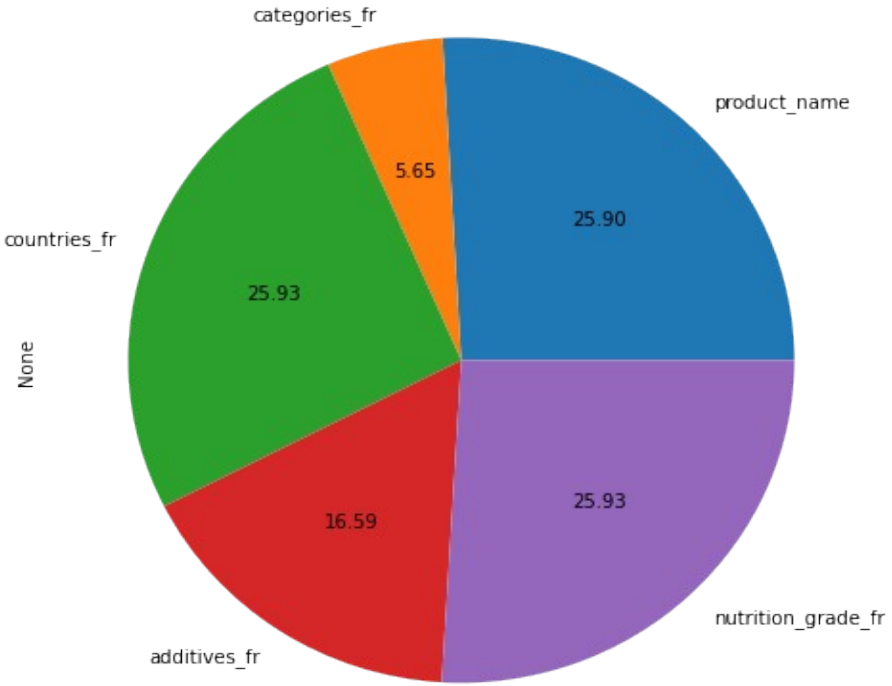
B-BOXPLOT


100
80
60
40
20
0
fat_100g
saturated-fat_100g
trans-fat_100g
carbohydrates_100g
fiber_100g
proteins_100g
salt_100g

REMARQUE BOXPLOT

-Hormis la variable 'carbohydrates_100g', les autres variables présentent un nombre d'outliers important.
-IQR tres important pour la variable 'carbohydrates_100g'; tres faibles pour les autres variables.
-Moyennes très tassées en dehors de la variable 'carbohydrates_100g'

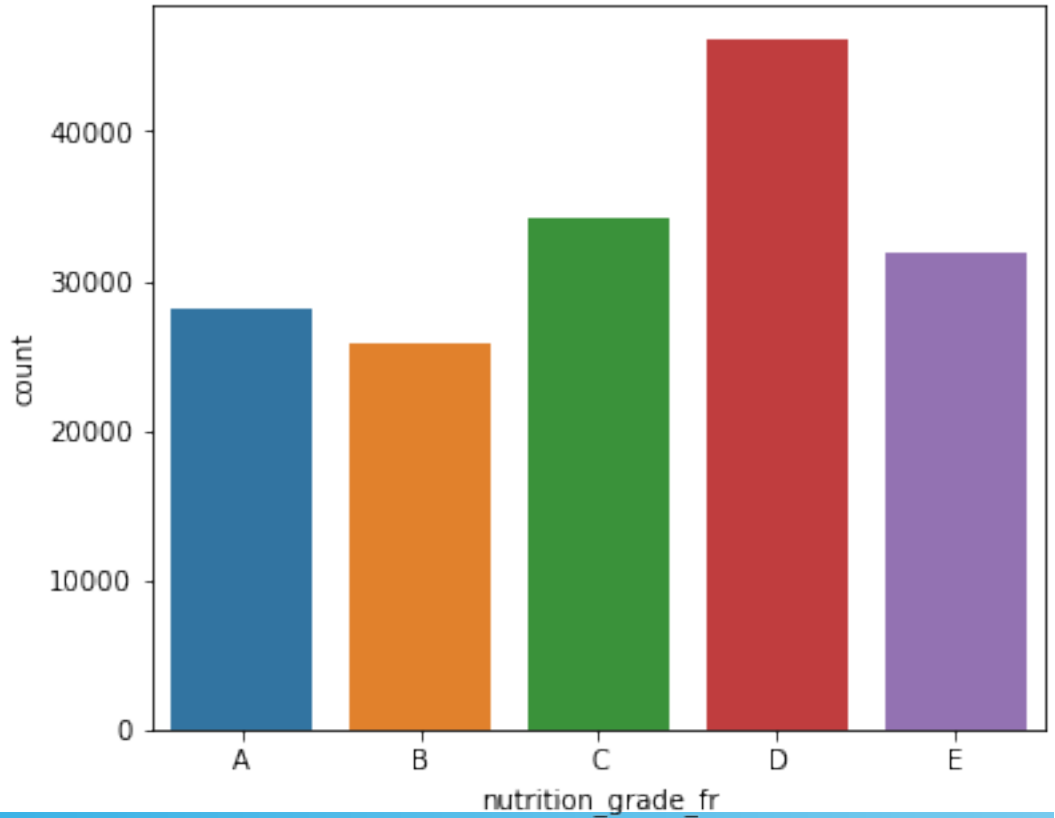
POURCENTAGE DE REMPLISSAGE DE VARIABLES



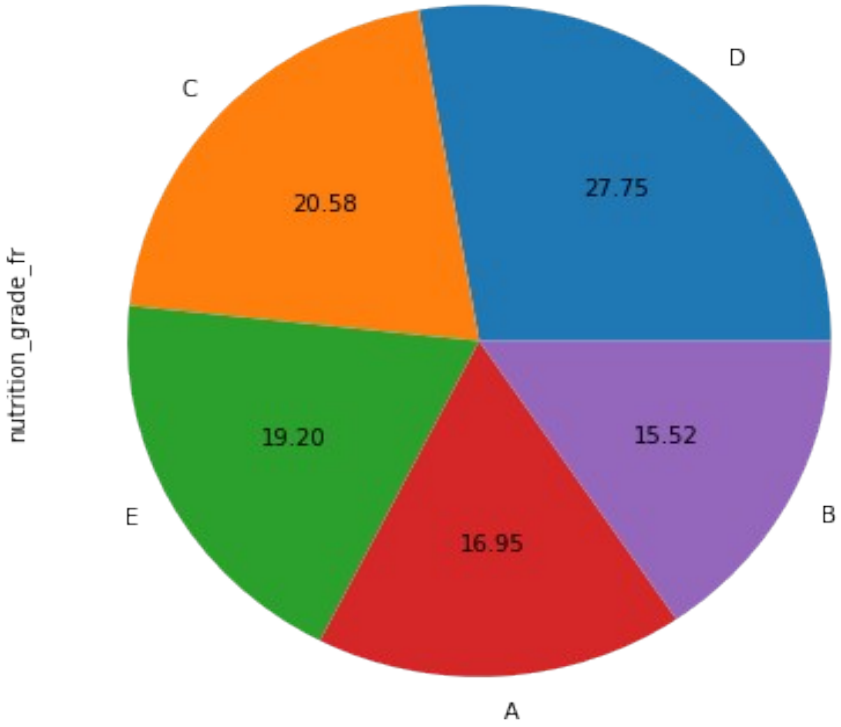
La variable catégorie présent un pourcentage

NUTRIGRADE REPARTITION:

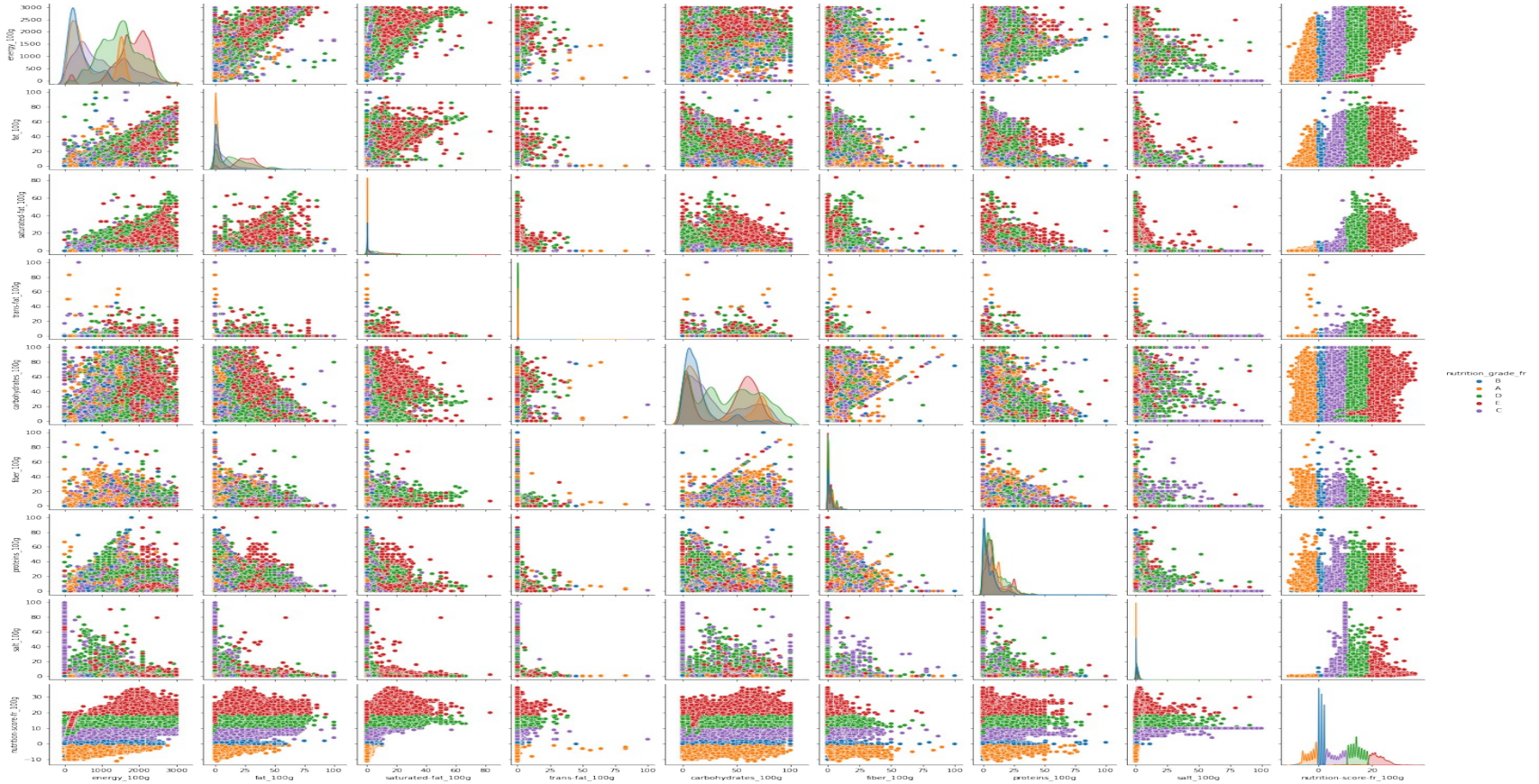
D 46141
C 34212
E 31923
A 28189
B 25813
Name: nutrition_grade_fr, dtype: Int64



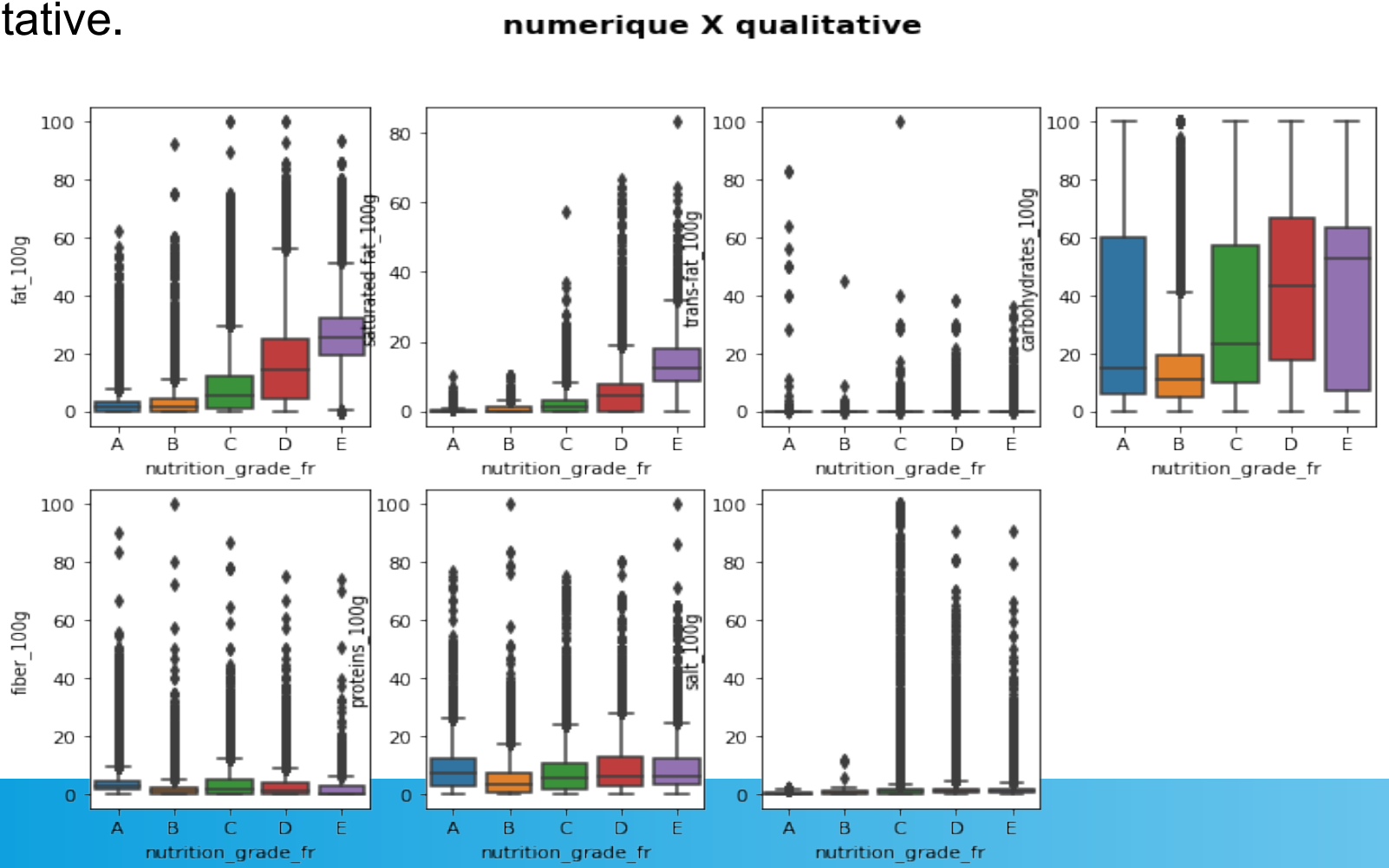
REPARTITION CLASSE NUTRIGRADE EN %



ANALYSE BIVARIEE : scatterplot variables numériques



ANALYSE BIVARIEE : boxplot variable numérique croisée avec une variable quantitative.



ANALYSE BIVARIEE : boxplot variable numérique croisée avec une variable quantitative.

REMARQUE

1-fat_100 ,satured_fat_100g,croissent en même temps que nutrition_grade_fr 'croit'. Relation à confirmer par une analyse ANOVA

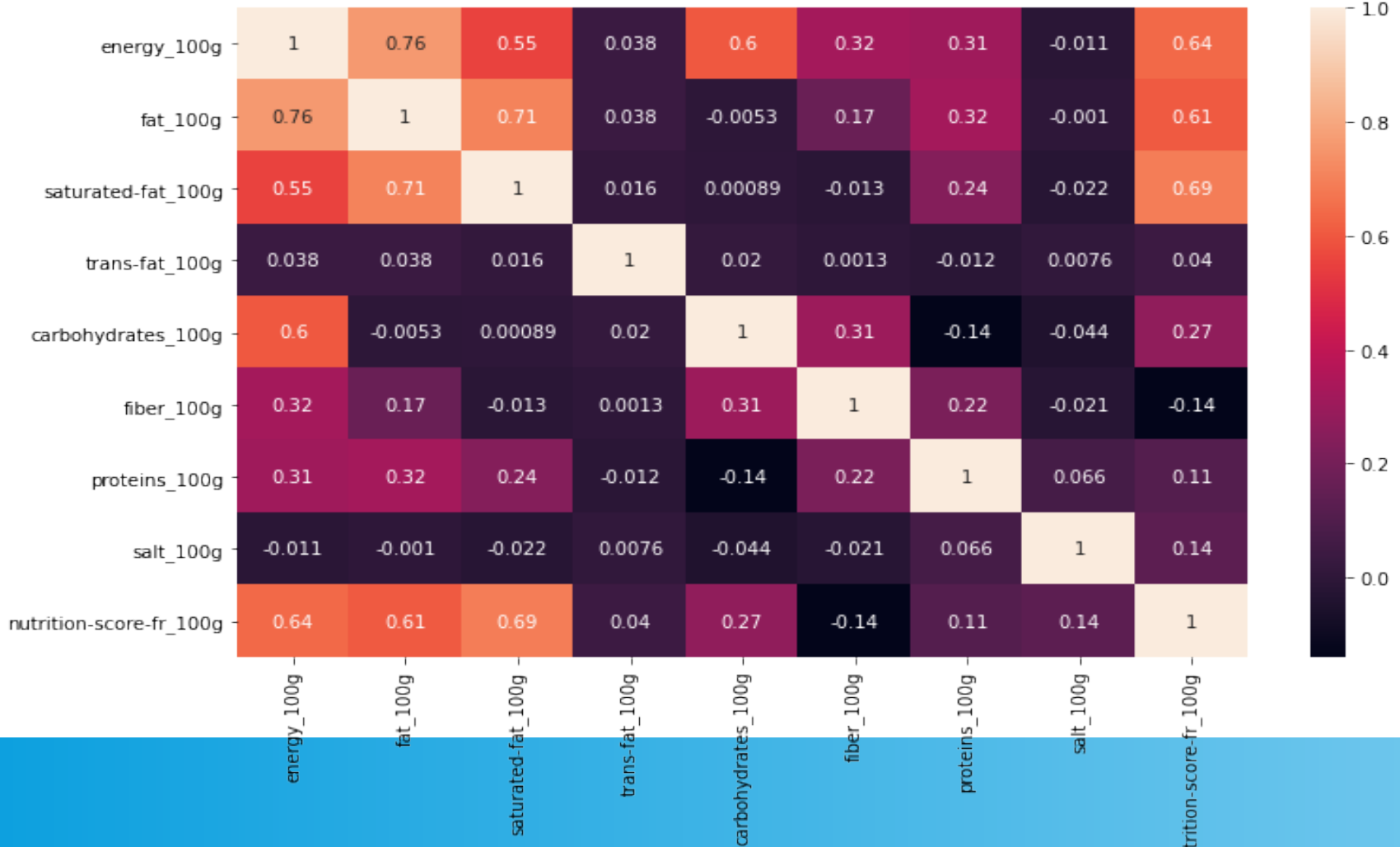
2-salt_100 croit légèrement en même temps que nutrition_grade_fr 'croit'
Pour les autres graphiques on ne peut conclure.

ANALYSE BIVARIEE : scatterplot variables numériques

REMARQUES :

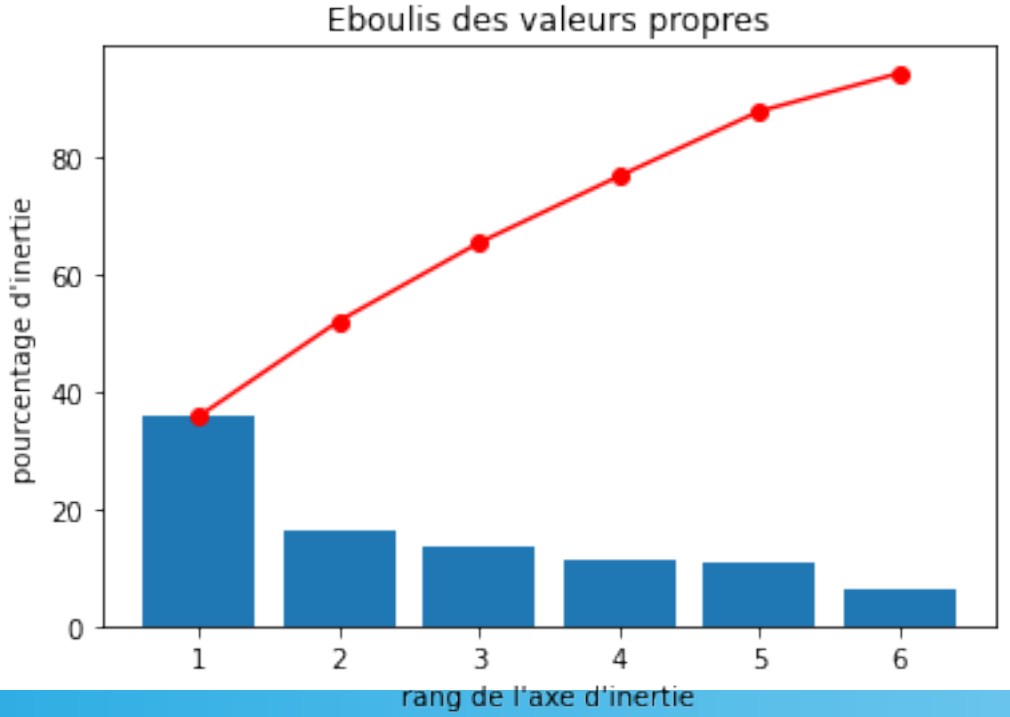
Difficile de cerner des relations. Classes très entremelées. On peut noter à l’ extrême droite la séparation plus ou moins nettes des classes pour chaque nutriment.

CORRELATION

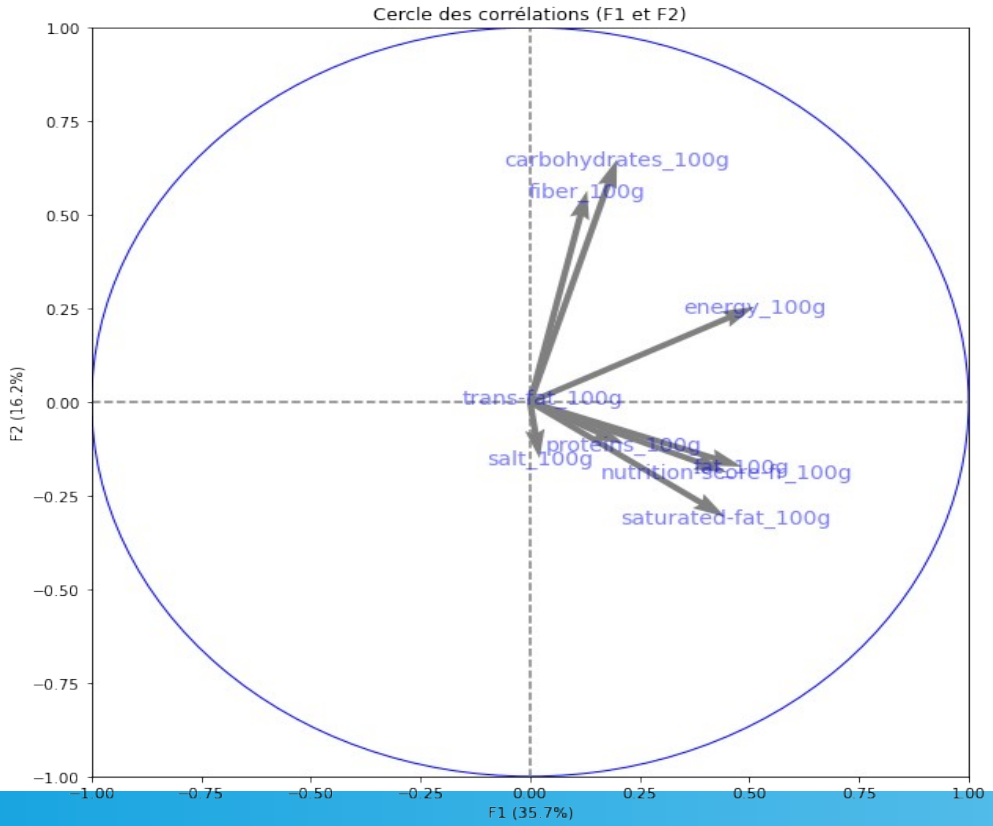


- **4-Analyse multivariée et résultats statistiques associés, en lien avec l’idée d’application**

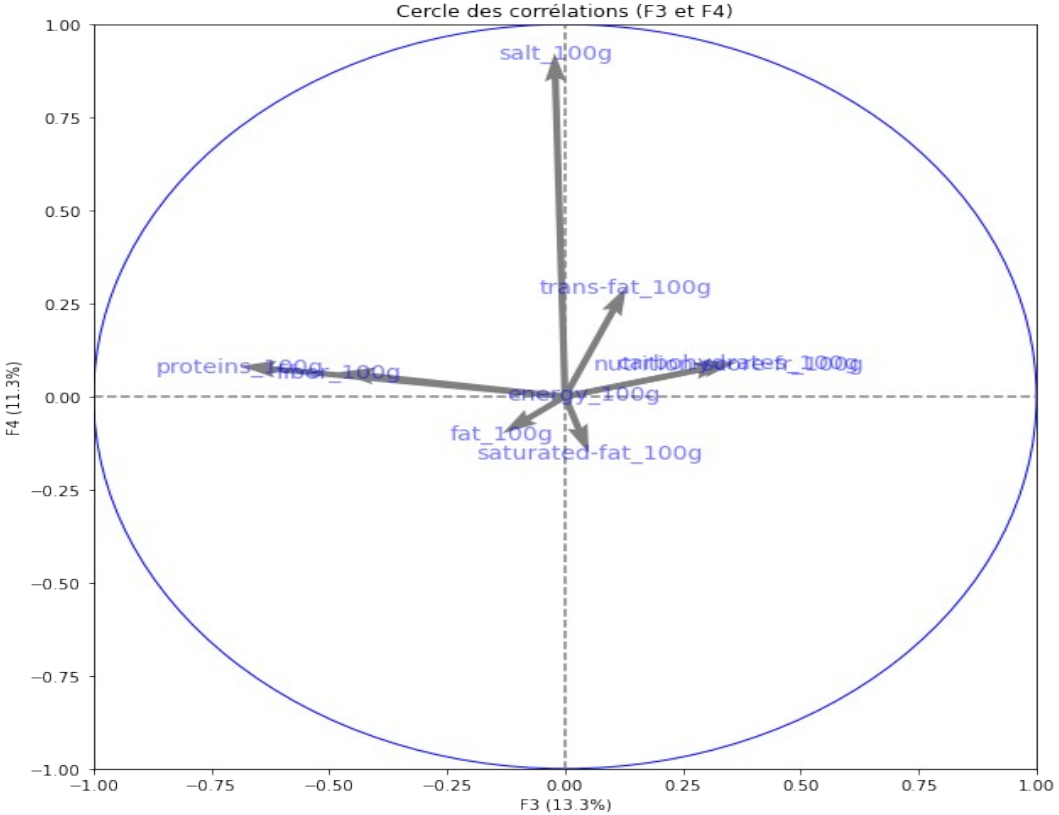
- **PCA : éboulis**
  - **En applicant la méthode du coude nous ne retenons que 3 facteurs**



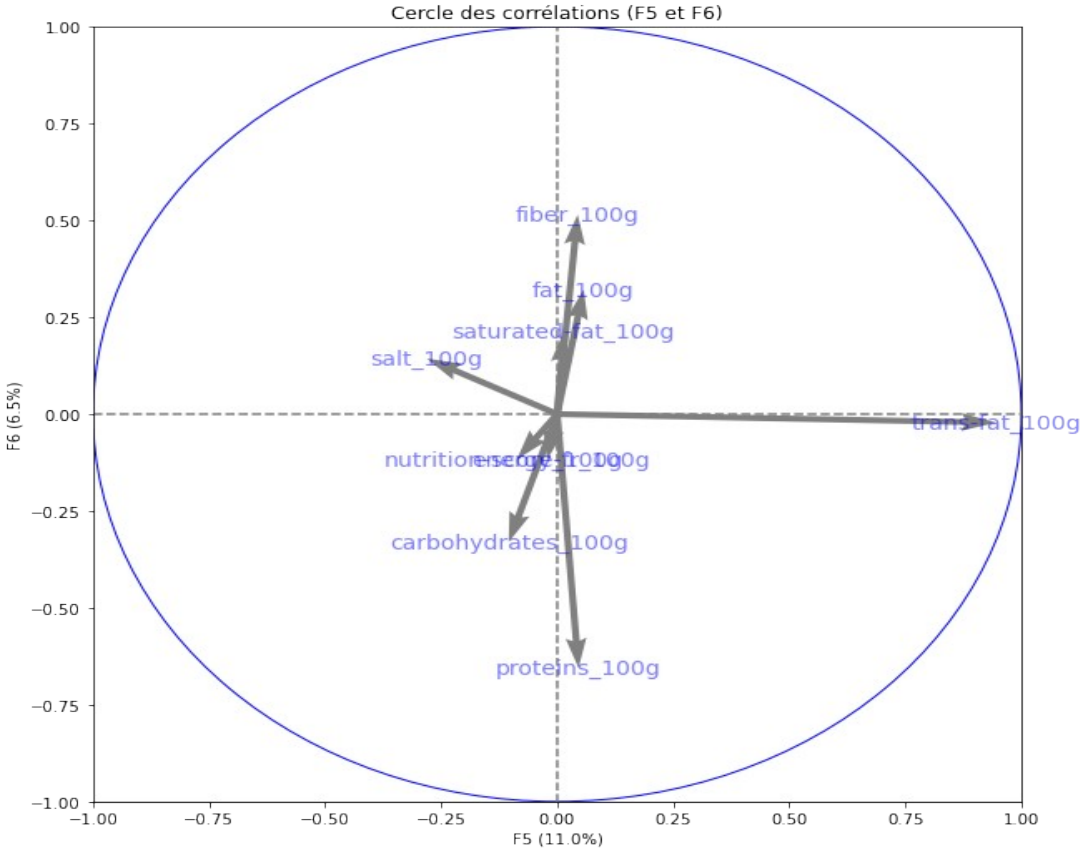
- **PCA : cercles de correlation axes F1 et F2.**



- **PCA : cercles de correlation axes F3 et F4.**



- **PCA : cercles de correlation axes F5 et F6.**

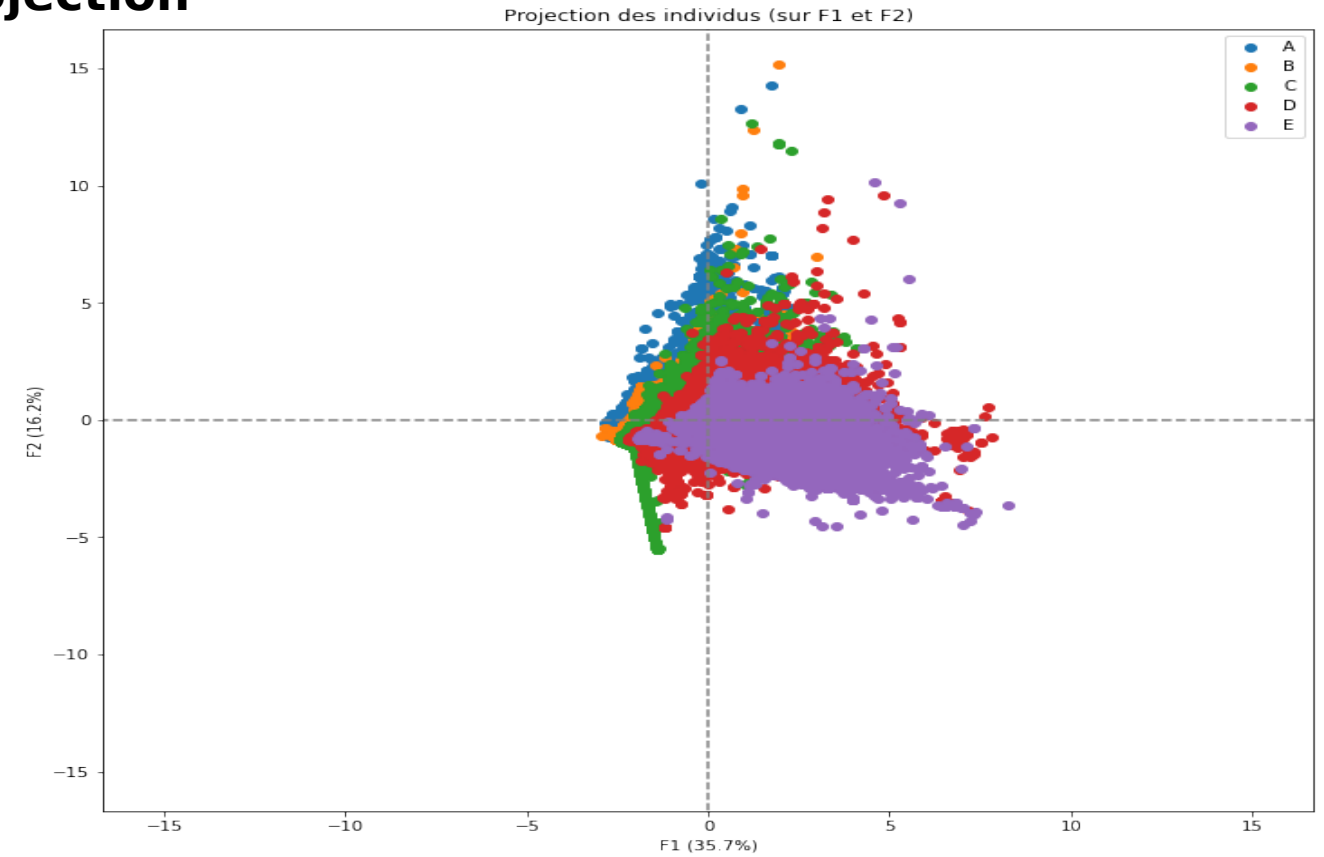


- **PCA : cercles de correlation**
- **REMARQUES**
-
- **-1er cercle (F1;F2):**
-
- Toutes les variables sont bien représentées en dehors des variables salt_100g et proteine_100
-
- On distingue 3 groupes de variables:
-
- a/ le groupe de variables carbohydrate_100g , fiber_100g très correlées entre elles et avec l'axe F2
-
- b/ energy_100 tres correlé avec l'axe F1 forme un groupe à une variable.
-
- c/ le groupe de variables satured_fat_100,nutrition_score_fr,fat_100g tres correlé avec l'axe F1 et entre elles.

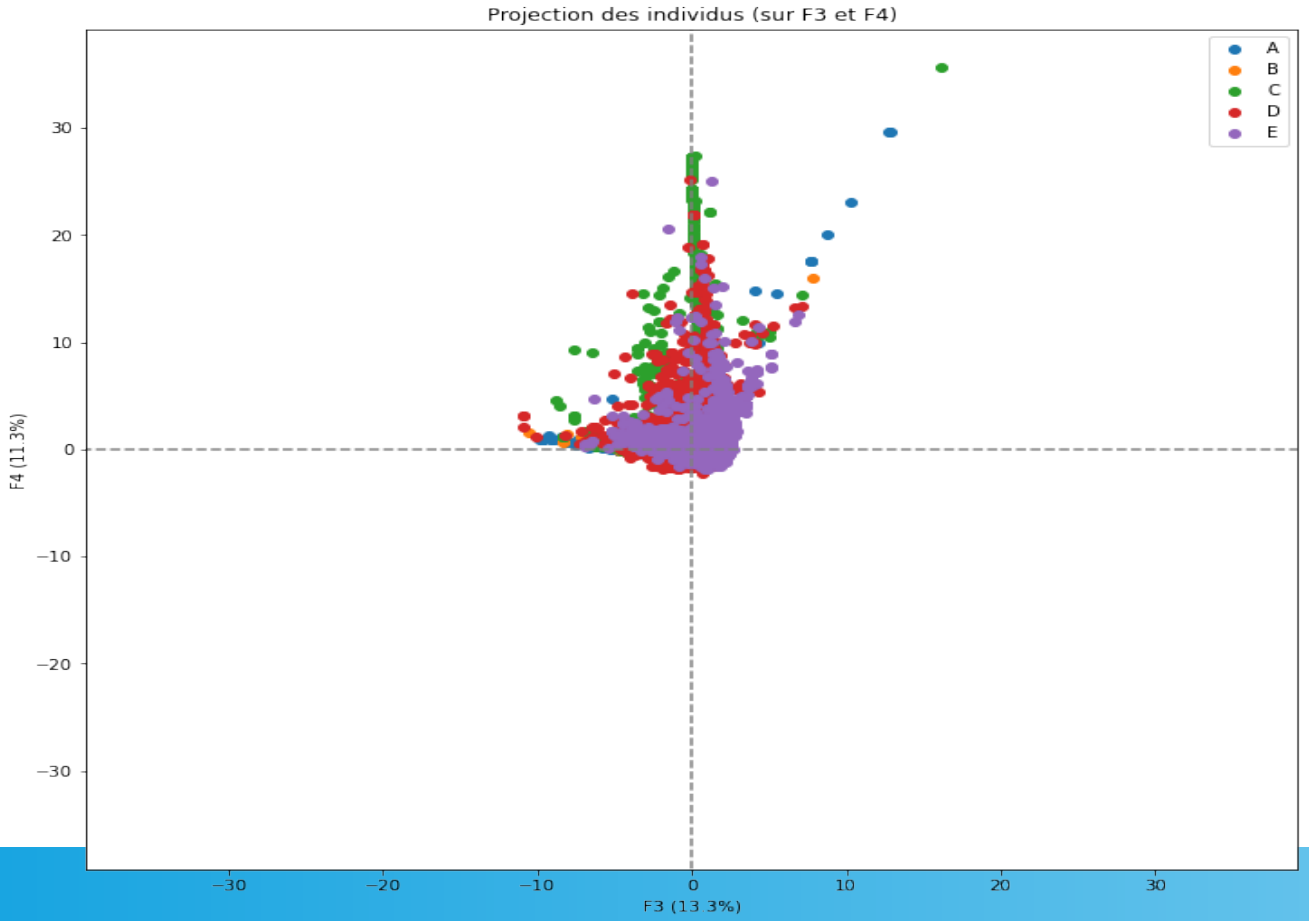
- **PCA : cercles de correlation**
- **REMARQUES**
-
- **-2e cercle (F3;F4):**
-
- Sont bien représentées le groupe de variables fiber_100g et protein_100,
- Ainsi que celui des variables nutrigrade_100 et carbohydrate_100.
- Les variables de ces 2 groupes sont très correlées avec l’axe F3
- Tandis la variable salt_100g très bien représentée est très correlée avec F4

- **PCA : cercles de correlation**
- **REMARQUES**
-
- **-3e cercle (F5;F6):**
-
- Sont bien représentées les variables fiber_100g et protein_100, et sont en
- Même temps en opposition sur laxe F6.
- La variables transfat_100 est bien représentée sur l’axe F5.

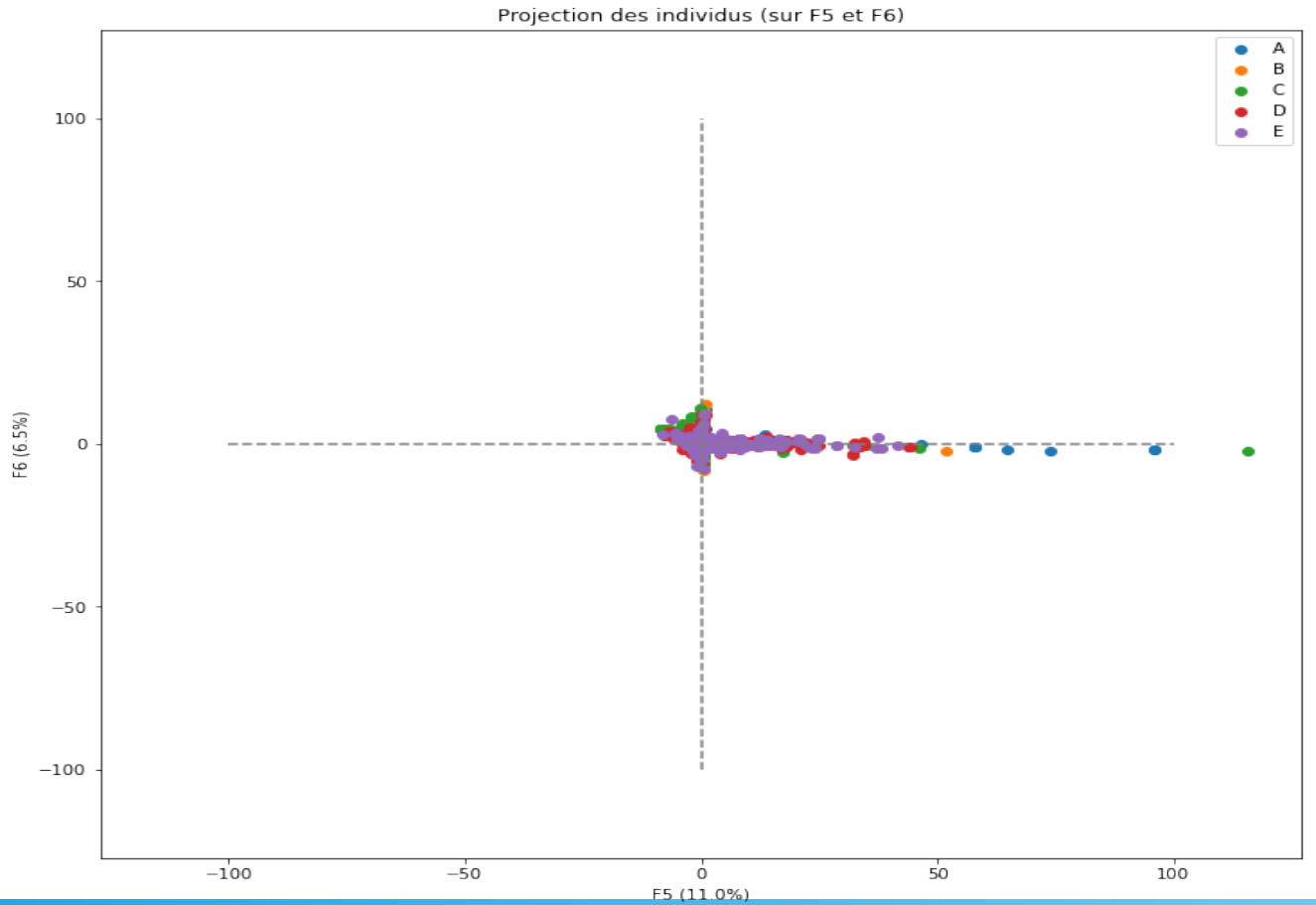
- **PCA: projection**
-



- **PCA: projection**
-



- **PCA: projection**
-



- **PCA: projection**

- **Classes très entremelées dans les graphiques de projection des individus.**

- **IDEE D’APPLICATION:**
-
- Nouvelle formule de calcul des points :

  Pour calculer le nouveau score des produits pour diabétiques

  Nous avons besoin de savoir

  -le nombre d'additifs rentrant dans la composition des aliments

  -l'impact tres négatif de certains d'entre eux comme:

  Propionate de calcium :E282 (Cet additif alimentaire pouvant déclencher le diabète et l'obésité)

  La Saccharine:E954 est aussi deconseillé.

  Pour corriger le nombre de points nutrition-score-fr_100g en fonction des éléments ci-dessus nous utiliserons la formule suivante:

  nombre total de points = nutrition-score-fr_100g + points_additifs

  1/ nutrition-score-fr_100g cette données est fournit dans le dataset

  2/ points_additifs est fonction du nombre d'additifs rentrant dans la composition d'un produit. A determiner.

Pour les produits contenant le E282 et E984 le nombre de points sera le nombre maximum.

-

- **APPLICATION: attribution de oints**

Points Produits(Hors Boissons)

-15 à -1_____Ad

0 à 2_____Bd

3 à 10_____Cd

11 à 18_____Dd

19 à 40_____Ed

Points Boissons

Eau_____Ad

<-1_____Bd

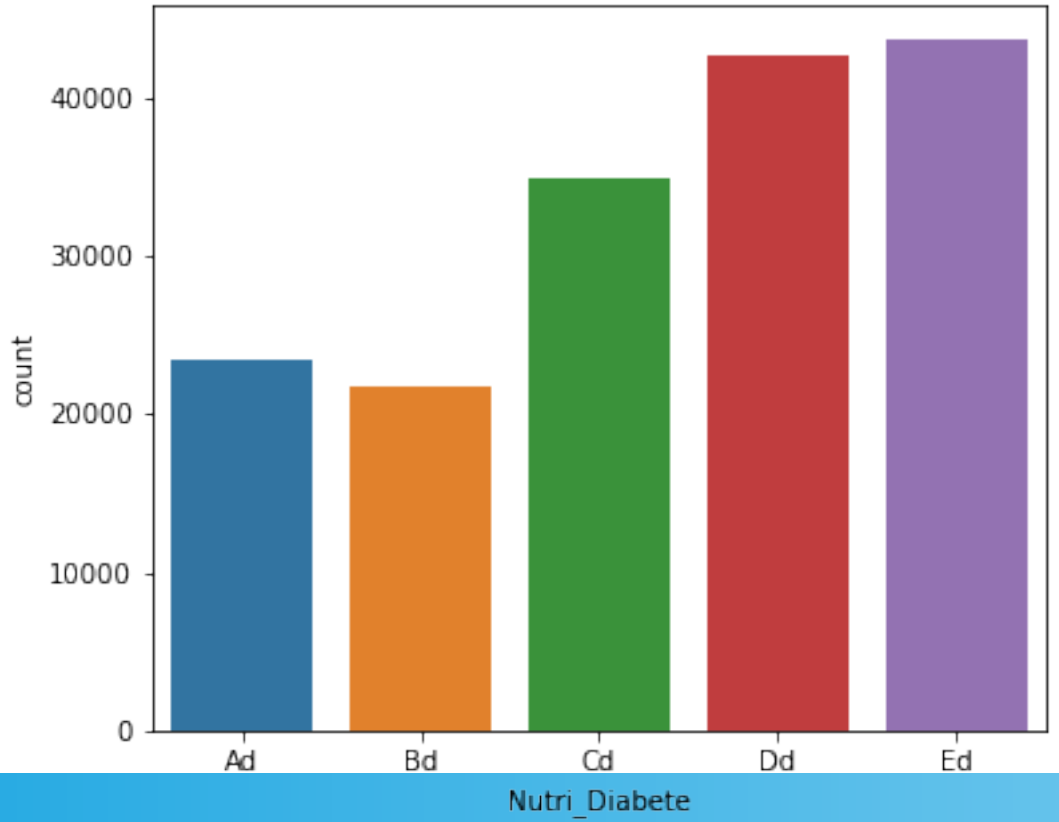
2 à 5_____Cd

6 à 9_____Dd

10 à 40_____Ed

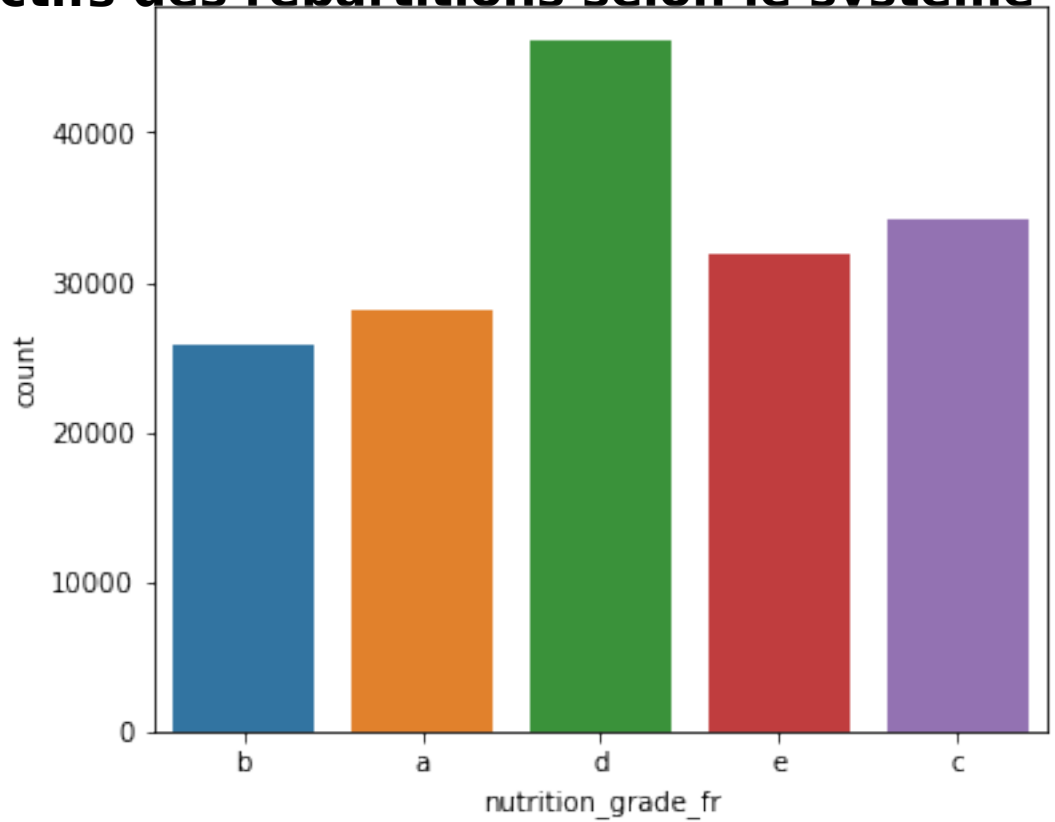
APPLICATION: effectifs des repartitions selon Nutri_Diabete

	Nutri_Diabete
Ad	23406
Bd	21728
Cd	34824
Dd	42665
Ed	43655



APPLICATION: effectifs des repartitions selon le système Nutri_Score

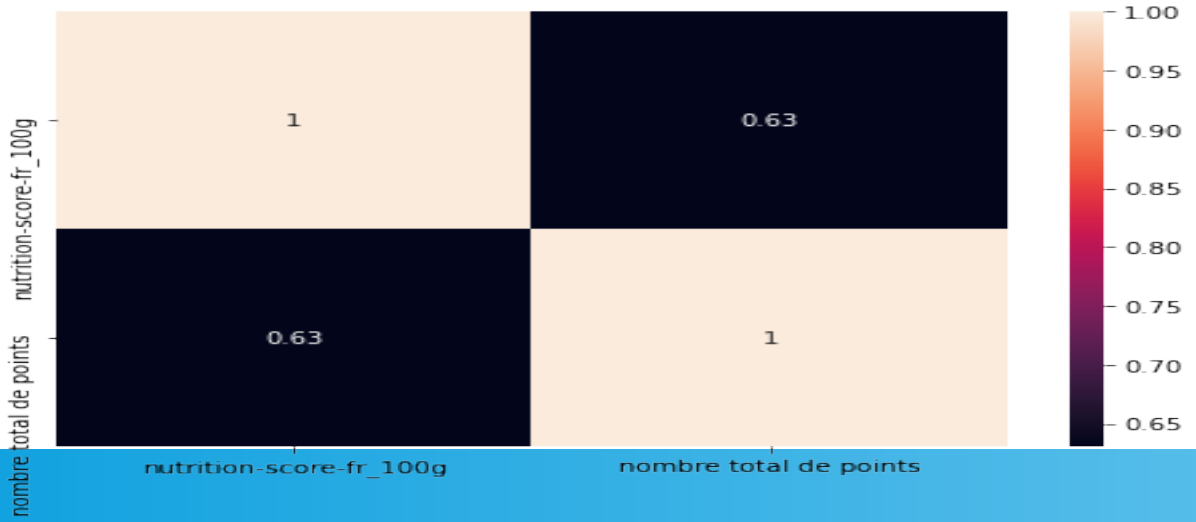
nutrition_grade_fr	
a	28189
b	25813
c	34212
d	46141
E	31923



APPLICATION: comparaison des 2systemes Nutri_diabete et Nutri_Score

CORRELATION :

corr=df200[['nutrition-score-fr_100g','nombre total de points']].corr()
corr



APPLICATION: comparaison des 2 systemes Nutri_diabete et Nutri_Score
CrossTab :

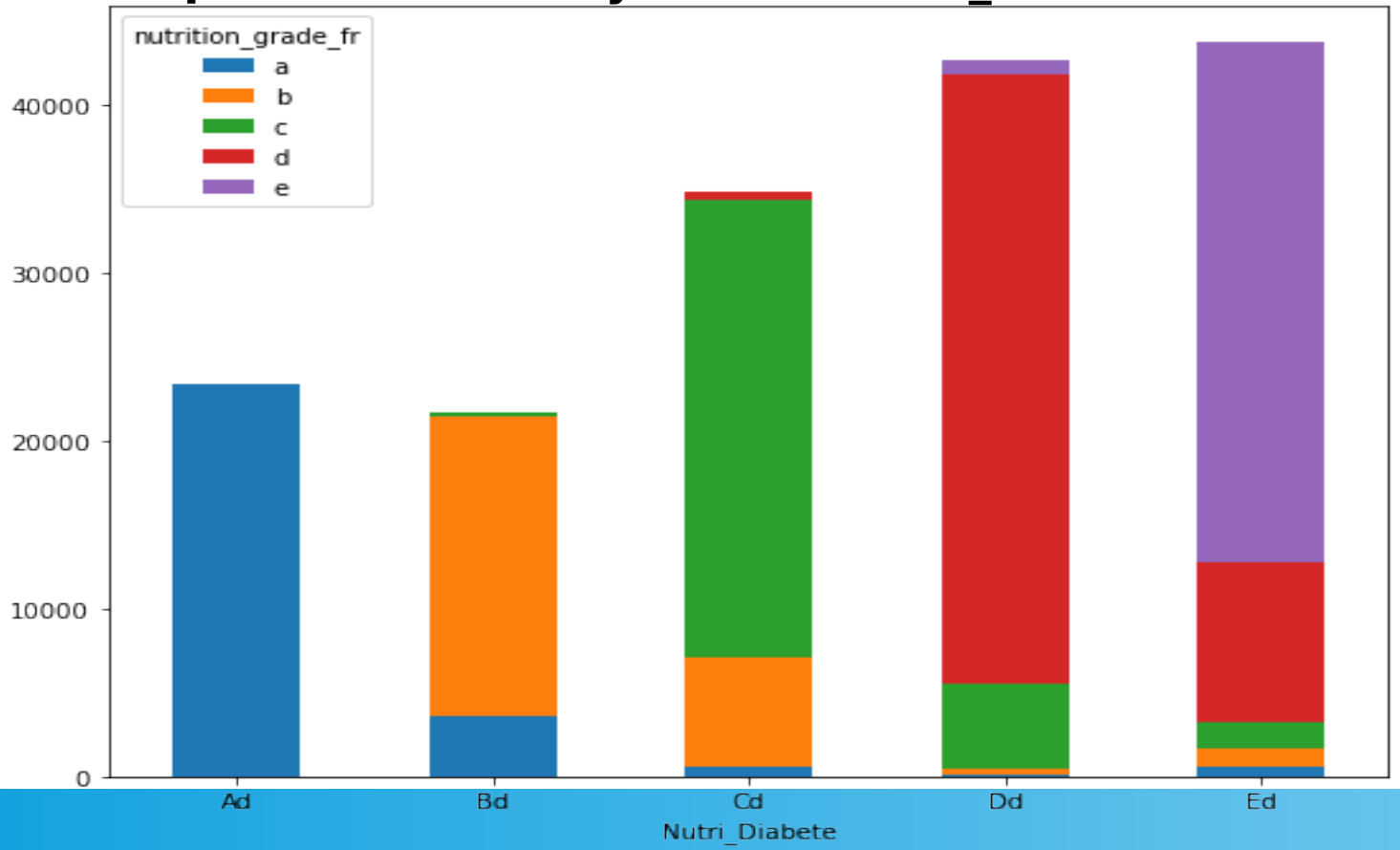
```
# TABLEAU CROISE ENTRE 'nutrition_grade_fr' ET 'Nutri_Diabete'
x=df200['nutrition_grade_fr']
y=df200['Nutri_Diabete']
cross=pd.crosstab(x, y, dropna=False,margins=False,normalize=False)
cross
```

APPLICATION: comparaison des 2 systemes Nutri_diabete et Nutri_Score
CrossTab :

```
# REPRESENTATION GRAPHIQUE CROSSTAB.
cross.T.plot(kind='bar',stacked=True, rot=0,figsize=(9,7))
```

APPLICATION: comparaison des 2 systemes Nutri_diabete et Nutri_Score

CrossTab :



APPLICATION: comparaison des 2 systemes Nutri_diabete et Nutri_Score
REMARQUE

Le tableau montre de façon pertinente que Nutri_Diabete surclasse nutriscore.

En examinant le graphique on remarque que:

1-Des produits classés 'a' se retrouvent répartis dans 3 classes de Nutri_Diabete: Ad,Bd,Cd.

2-Des produits classés 'b' se retrouvent répartis dans 4 classes de Nutri_Diabete: Bd,Cd,Dd,Ed.

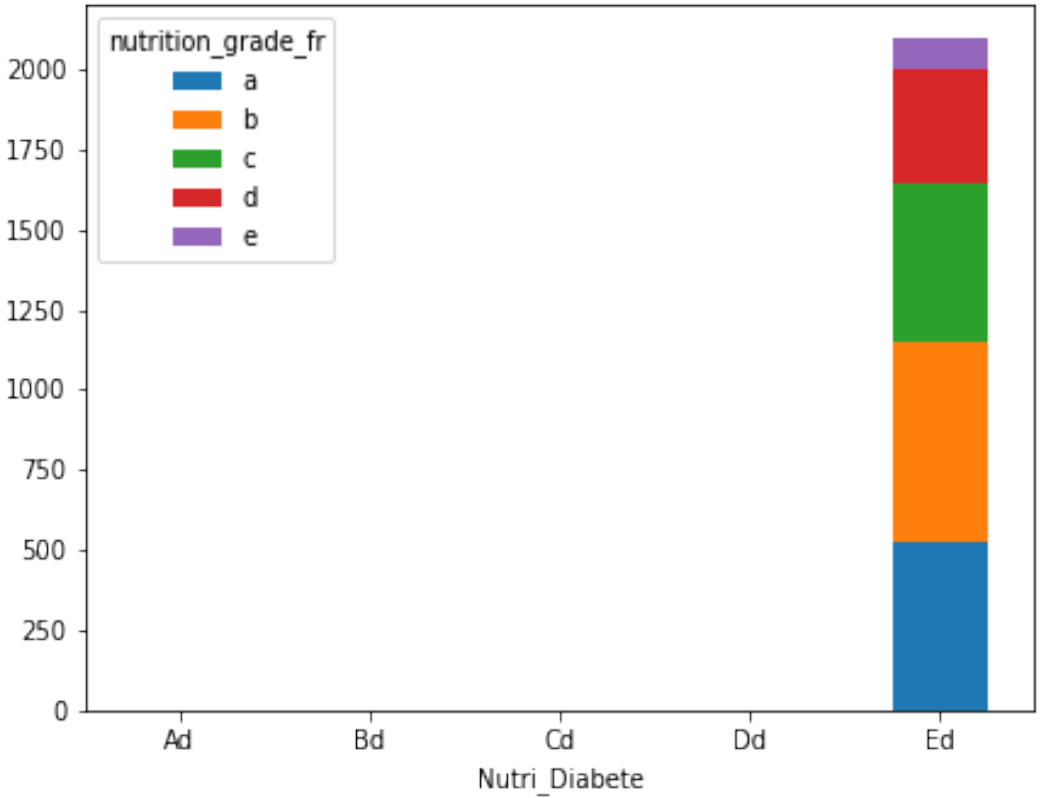
3-Des produits classés 'c' se retrouvent répartis dans 3 classes de Nutri_Diabete: Cd,Dd,Ed.

4-Des produits classés 'd' se retrouvent répartis dans 2 classes de Nutri_Diabete: Dd,Ed.

Dans l'ensemble d' après les résultats ci-dessus Nutri_Diabete classe correctement les produits contrairement à nutriscore.

- Pour le système **NutriScore** nous Constatons que des produits déclassés (classe Ed) par le système **NutriDiabète**
- sont répartis dans plusieurs classes:

- 525 dans A
- 622 dans B
- 497 dans C
- 357 dans D
- 95 dans E.



5-CONCLUSION

ON note une corrélation moyenne entre les 2 systemes (graphique de correlation)

1-Nutriscore peut aider à sélectionner un certains nombre d'aliments "sains" mais le filtrage souffre encore de faiblesse.

2-Par ailleurs l'index glycémique des produits n'est pas renseigné.

3-autre contrainte pour une application pour les diabétiques: connaitre le taux de glycémie du malade à l instant t et son poids.

4-les effets des additifs sont mal connus en dehors du E282 et E954 dont les effets secondaires sont certains.

5-Le Nutri_Diabete proposé elimine plus de produits nocifs pour les diabétiques mais en tenant compte des remarques 2/ 3/ 4/ il ne saurait etre performant par insuffisance ou absence d' information.

6-La base de données open.food devrait être reconsidérée. La saisie des données peut être améliorée pour permettre une exploitation optimale.