Projet 2

Concevez une application
au service
de la santé publique

Application d’aide au diabétique

- Le diabète est une maladie chronique qui survient lorsque le pancréas ne produit pas suffisamment d’insuline . L’insuline est une hormone régulatrice de la glycémie(taux de glucides dans le sang) .
- Un diabétique en **hypoglycémie** à besoin de glucides simples(sucre, confiture…)
- Un diabétique ayant une **glycémie normale** doit privilégier les glucides complexes (féculent, céréales,,)
- Un diabétique en **hyperglycémie** doit privilégier les aliments pauvres en glucides (légumes)
- Nous aimerions développer une application pour aider les diabétiques à calculer les glucides nécessaires lors de leur repas .
- Nous essayerons de calculer un Gluciscore à partir de ces données (A pour riche en glucide à F faible en glucide), ainsi qu’un système de recommandation en fonction des besoins des malades du diabète

Glusiscore

- Proposer des aliments riches en glucide si le diabétique à un besoin urgent de glucide et lui permettre de calculer précisément l’apport en glucide de chaque aliments
- A contrario lui conseiller l’inverse si le diabétique doit éviter les glucides

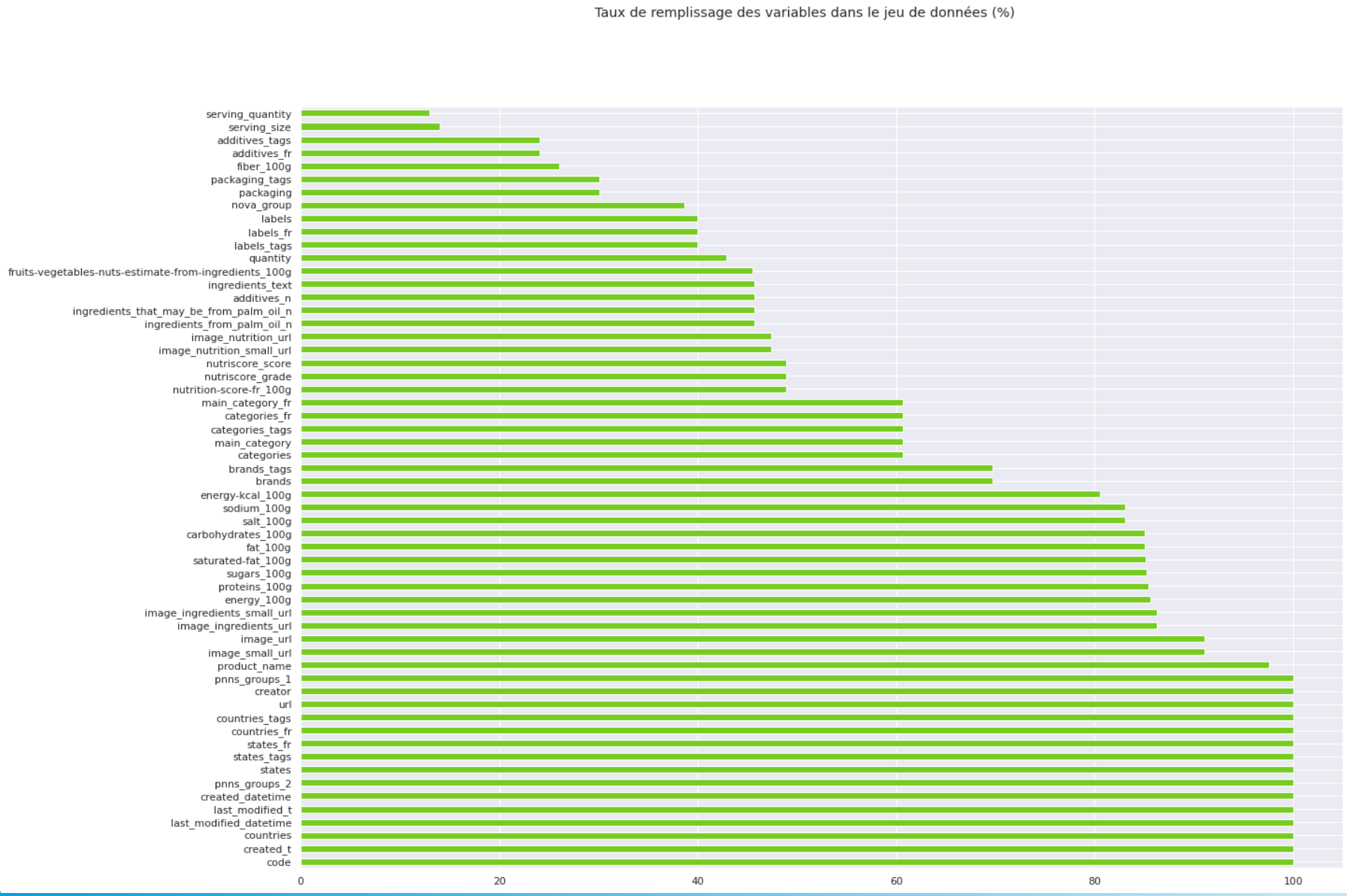
Source OpenFoodFacts

- 2051571 lignes 187 colonnes
- l’étude est suffisante sur la France avec 435465 lignes
- Focus particulier sur les informations relatifs aux glucides
- carbohydrates_100g et Fiber_100g
- Lien avec les autres valeurs nutritionnelles

Précision du jeux de données

- Informations générales sur la fiche du produit : nom, date de modification, etc.
- Un ensemble de tags : catégorie du produit, localisation, origine, etc.
- Les ingrédients composant les produits et leurs additifs éventuels.
- Des informations nutritionnelles : quantité en gramme d’un nutriment pour 100 grammes du produit.

Jeux de données

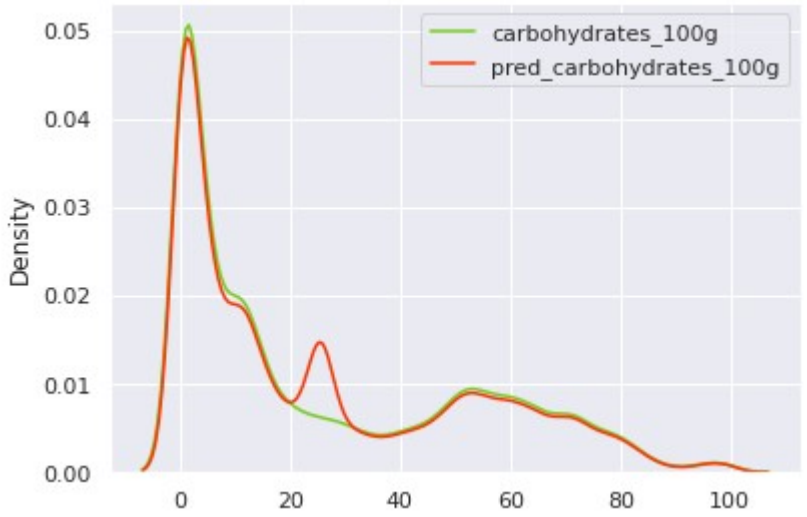


Valeurs manquantes

- Mise en forme des dates
- Suppression des doublons
- nettoyage des valeurs aberrantes sur les variables quantitatives et qualitatives
- Nettoyage sur l'énergie Kj kcal
- Sur les quantités exprimées par 100g (intervalle possible)
- Traitement des valeurs manquantes par knn :

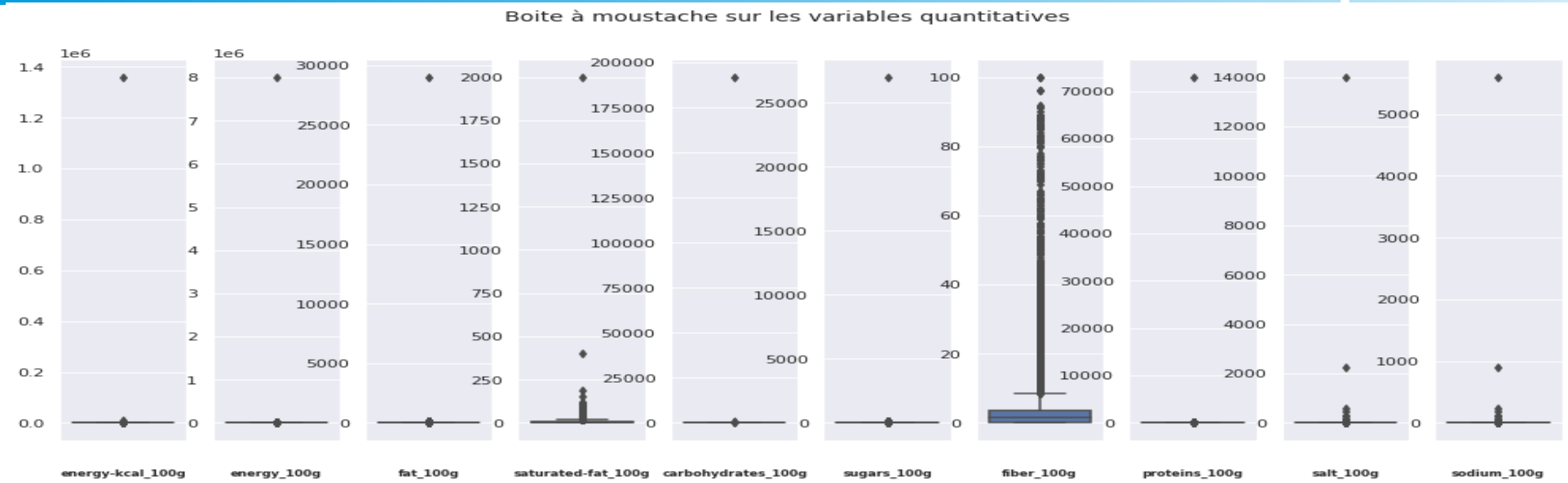
  variables imputer par knn

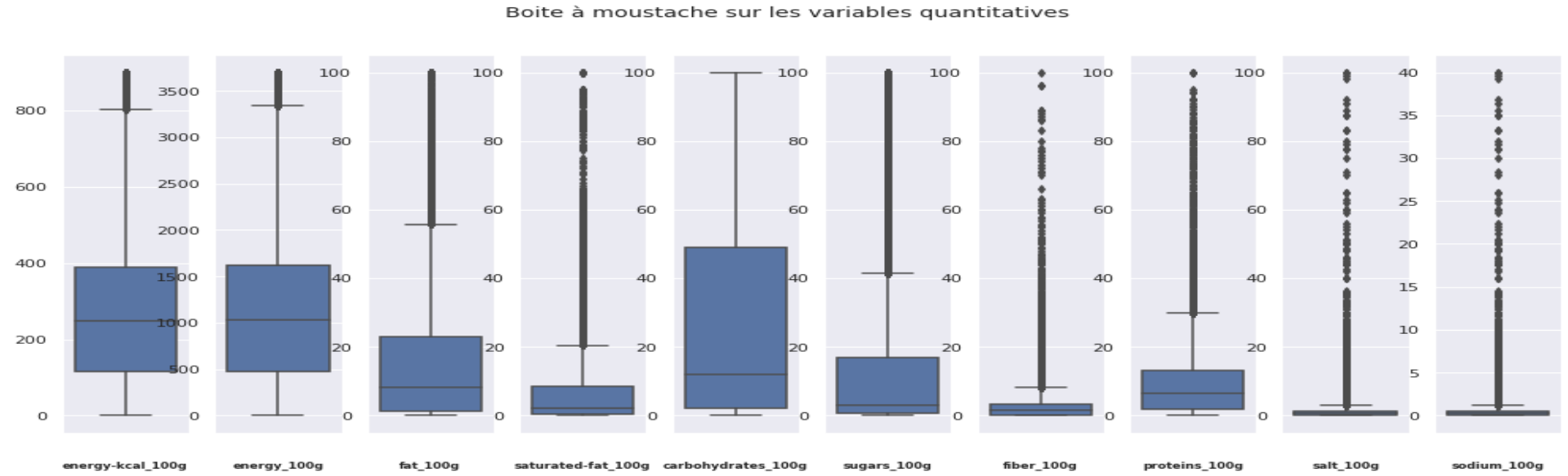
'energy_100g'	→ 13,88% de valeur manquantes
'proteins_100g'	→ 5.41 % de valeur manquantes
'saturated-fat_100g'	→ 5.91 % de valeur manquantes
'sugars_100g'	→ 5.75 % de valeur manquantes
'salt_100g',	→ 6,97 % de valeur manquantes
'carbohydrates_100g'	→ 5.33 % de valeur manquantes
'sodium_100g',	→ 6,97 % de valeur manquantes
'fat_100g'	→ 5.35 % de valeur manquantes



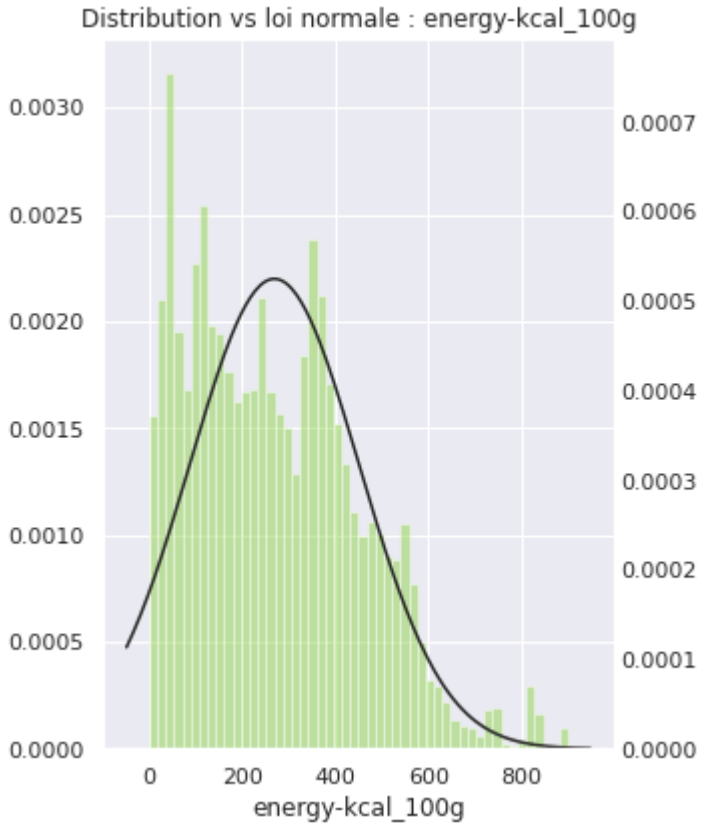
- Calcul des données fibres (fiber_100g) manquantes par la médiane car plus fiable
- Ils nous restent 25 colonnes qualitatives et 18 colonnes quantitative après nettoyage, 216566 lignes

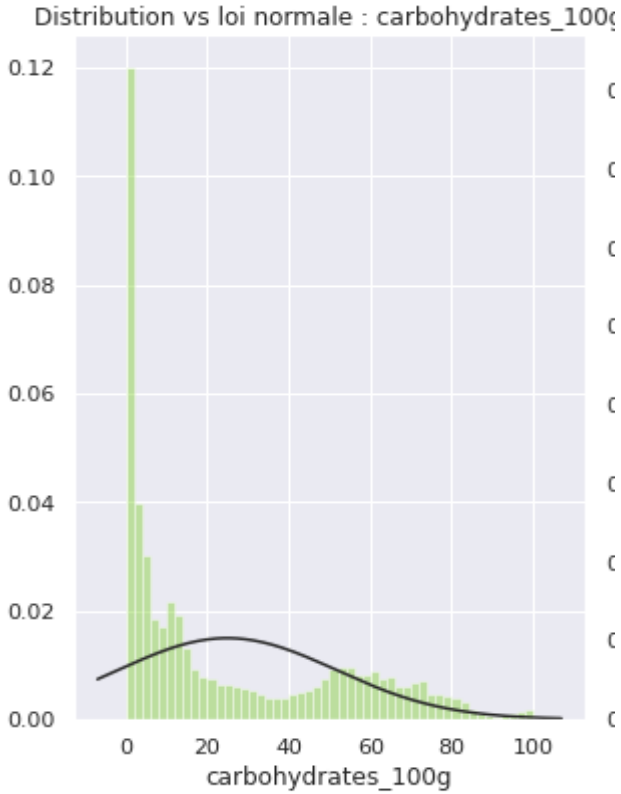
Valeurs aberrantes avant/après



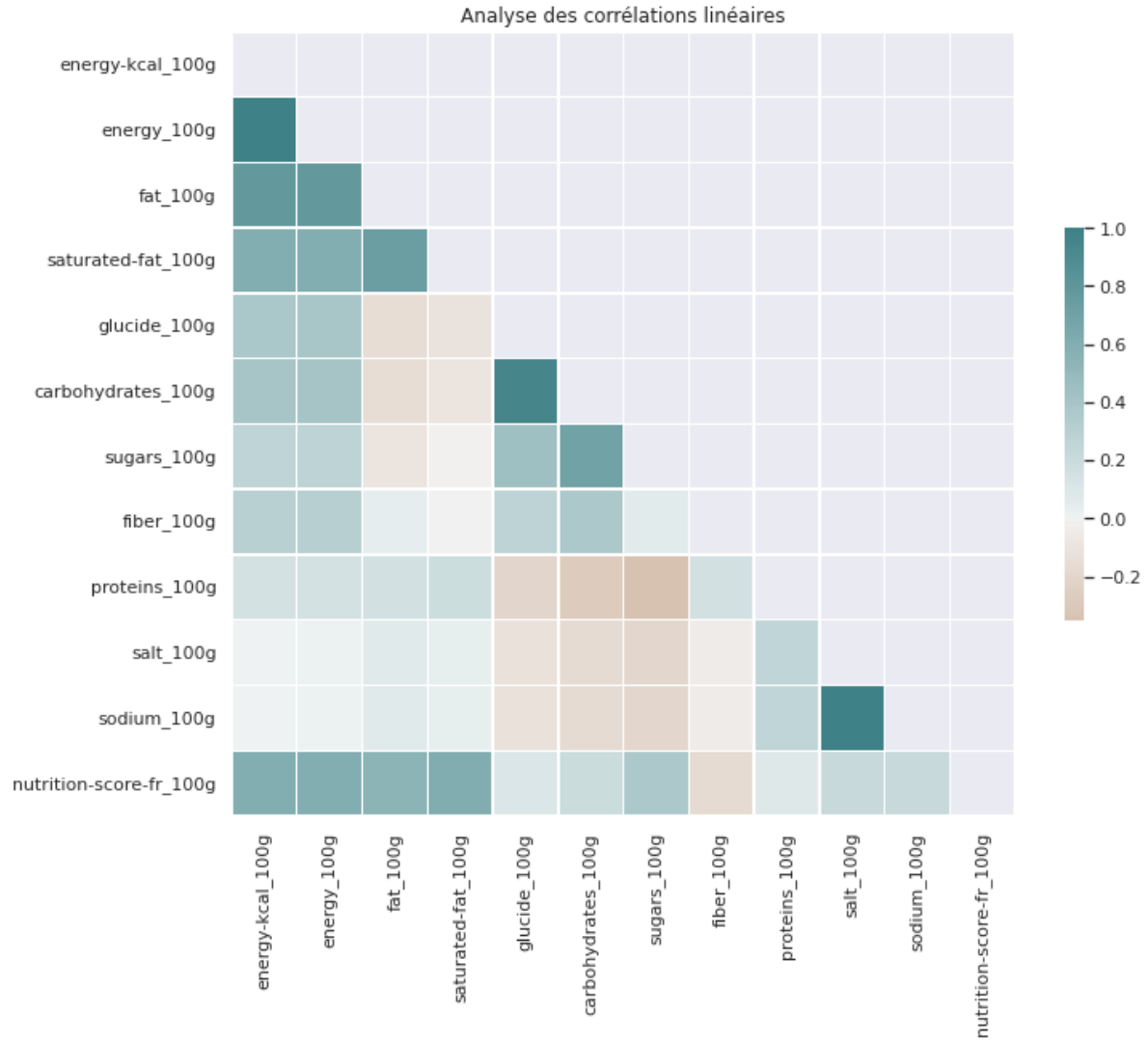


Analyse univariée





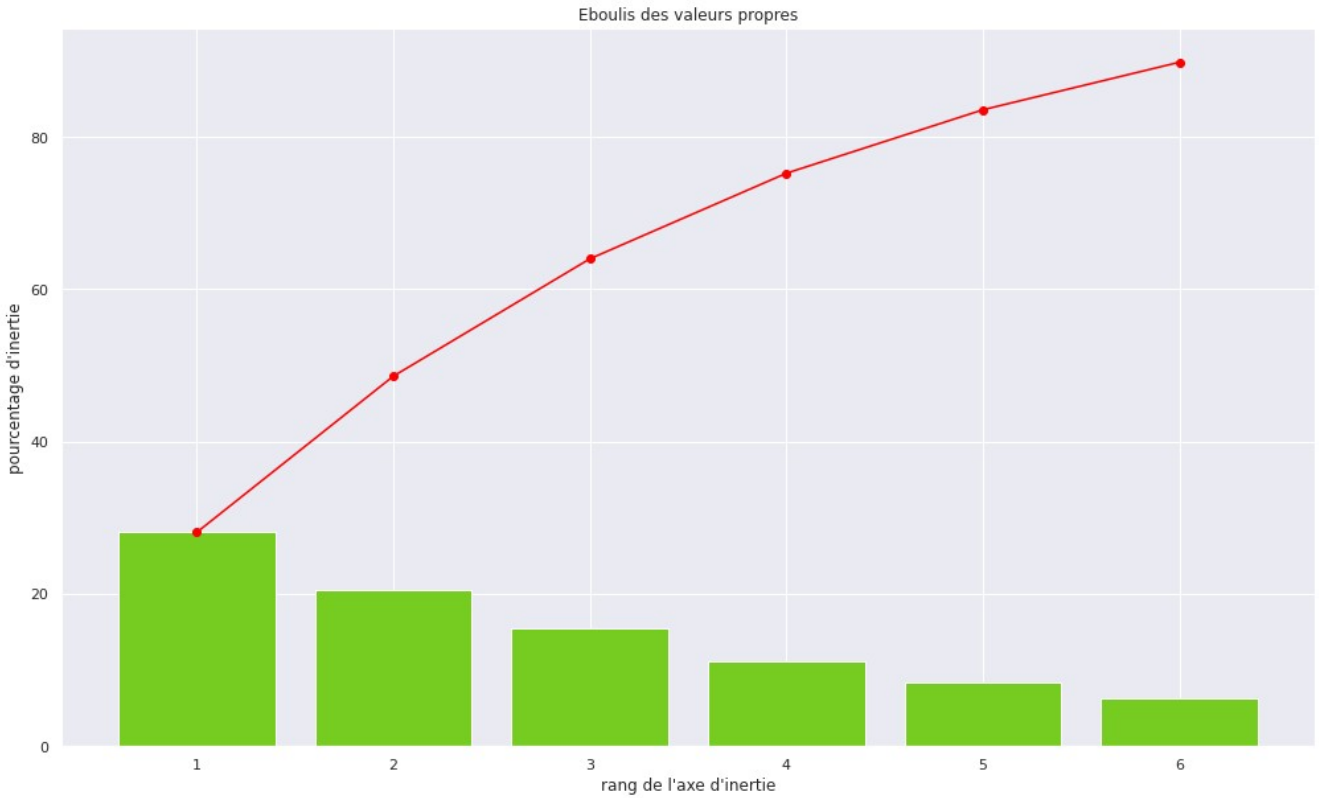
analyse des corrélations linéaires

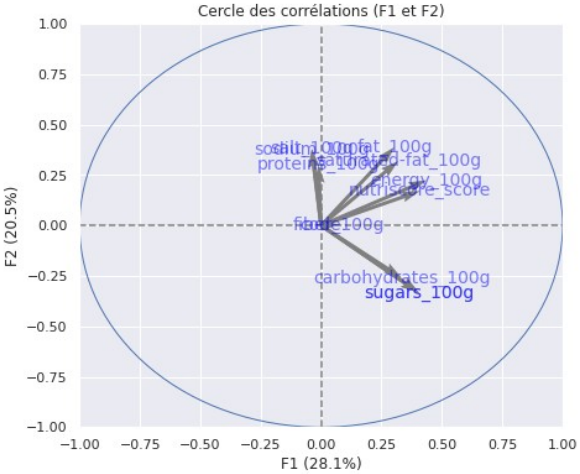


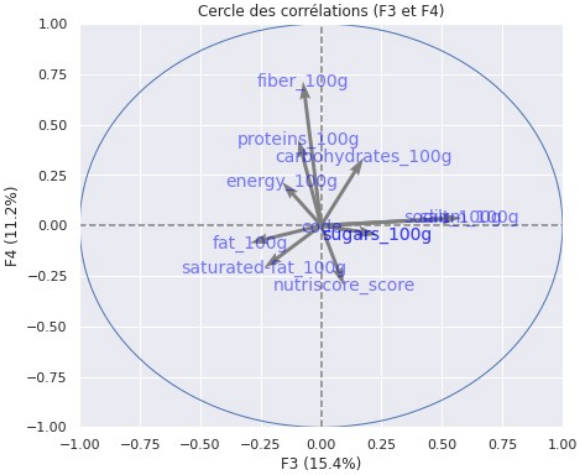
- **glucide** : corrélation avec energy
- **energy_100g** : corrélation avec fat_100g, saturated-fat_100g,carbohydrates_100g et nutrition-score-fr_100g
- **fat_100g** et **saturated-fat_100g** corrélés
- **sugars_100g** : corrélation avec carbohydrates_100g
- **sodium_100g** corrélé avec salt_100g
- **nutrition-score-fr_100g** : corrélation avec energy_100g ,energie-kcal_100g, saturated_fat_100g et fat_100g

analyse des composantes principales

Réduction à 6 ou 7 dimensions possible







Les glucides

Calcul des glucides :

Dans le calcul des glucides, il est recommandé de soustraire, du nombre total de glucides, seulement la moitié (50%) de la quantité de sucres-alcools déclarée sur l'étiquette, dans le tableau de valeur nutritive. Voici un exemple :

Un aliment contient 28 grammes (g) de glucides par portion, dont 4 g de maltitol et 2 g de fibres.
Donc, on calcule :
28 g – **(50 % de 4 g)** – 2 g = 28 – **(0,50 x 4)** – 2 = 28 – 2 – **2** = 24 g

Lorsqu'on applique le niveau avancé du calcul des glucides, on conclut que cette portion d'aliment contient 24 g de glucides.

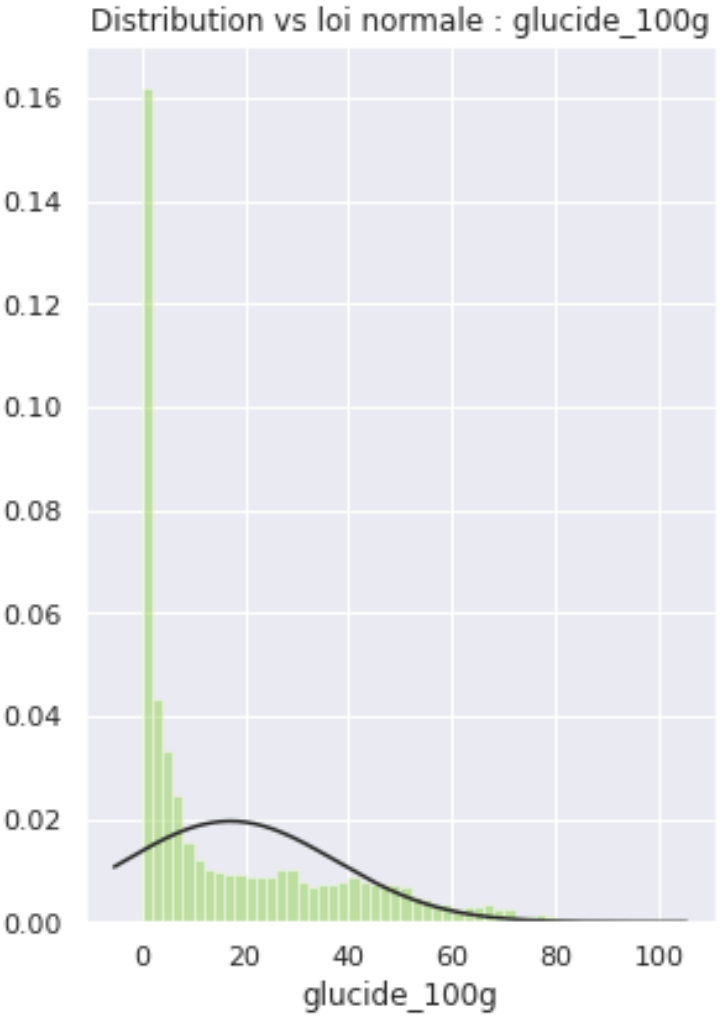
df['glucide_100g'] = df['carbohydrates_100g'].fillna(0) - (df['sugars_100g'].fillna(0)/2) – df['fiber_100g'].fillna(0)

Calcul du gluciscore :

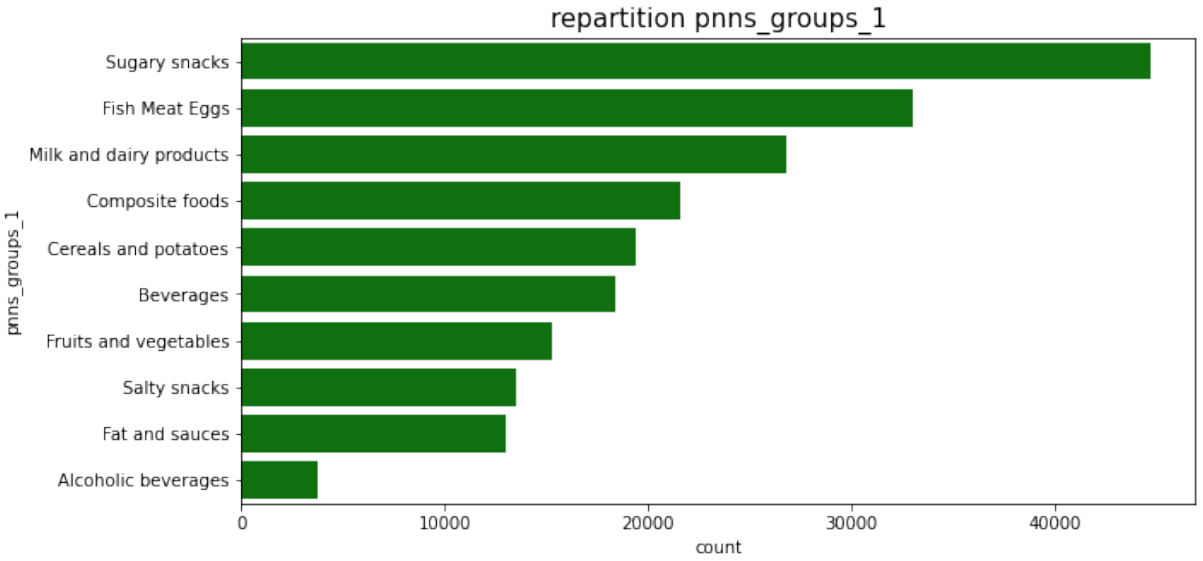
Nous calculerons ce score en fonction des intervalles de nutriments estimer bon ou pas pour chaque information nutritionnelles , en fonction de la catégorie de l aliments nous appliquerons un système de bonus malus

Zoom

Répartition des glucides



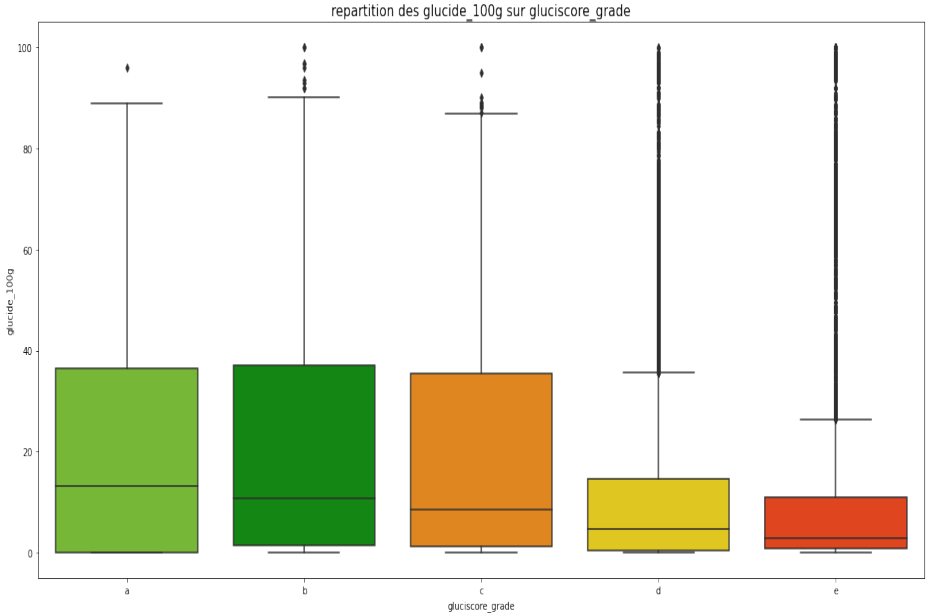
Répartition des groupes dans pnns_group1

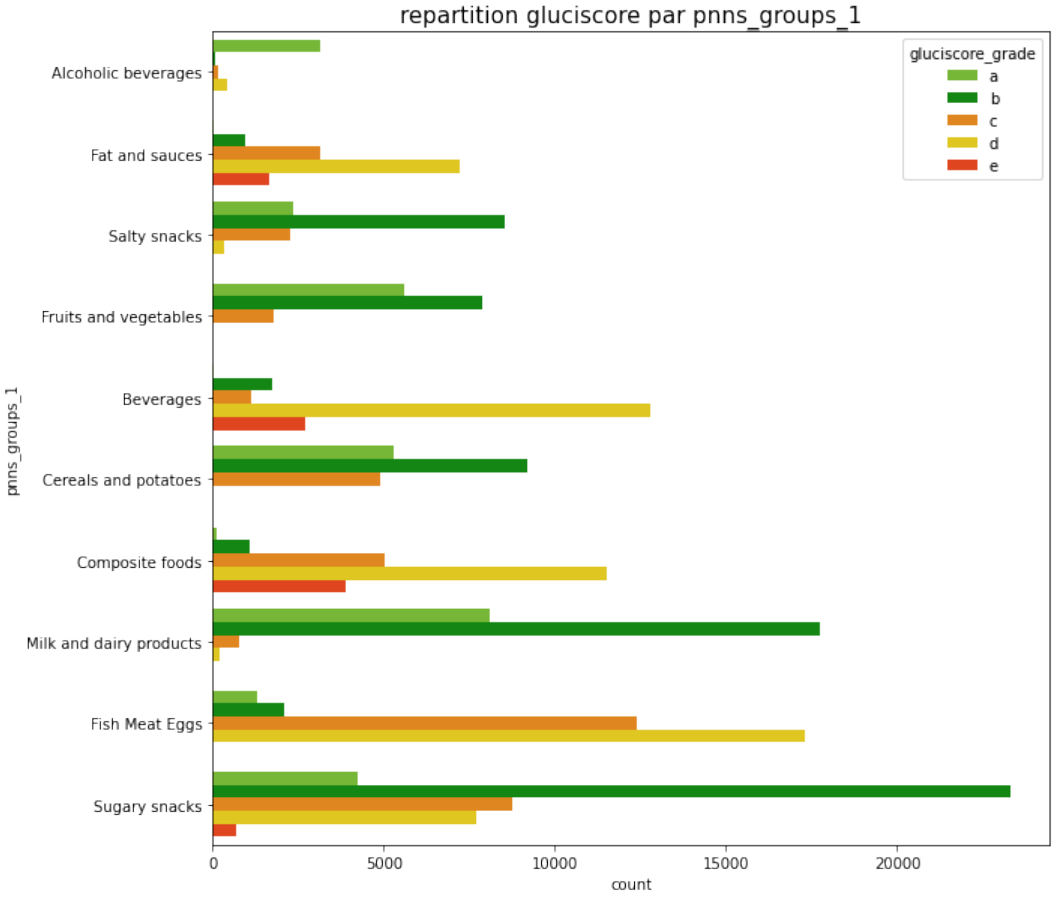


Analyse bivariée

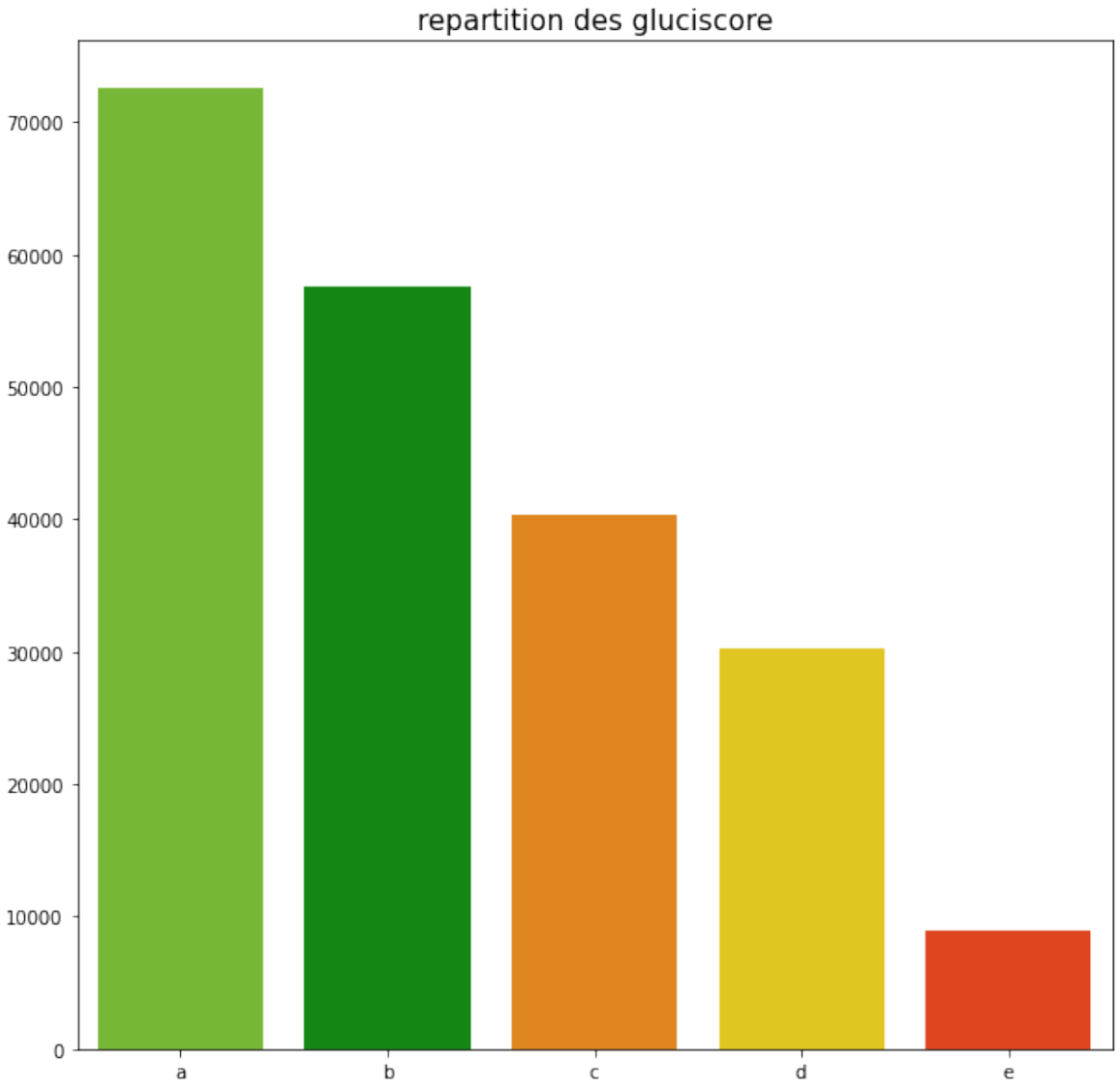
Calcul du Glusiscore

Nous allons raisonner en terme de poids , un aliment qui faire partie d'un groupe de catégorie connue comme la céréale aura une bon poids (score), alors qu'au contraire si le produit fait partie d une catégorie connue comme mauvaise le score sera petit.

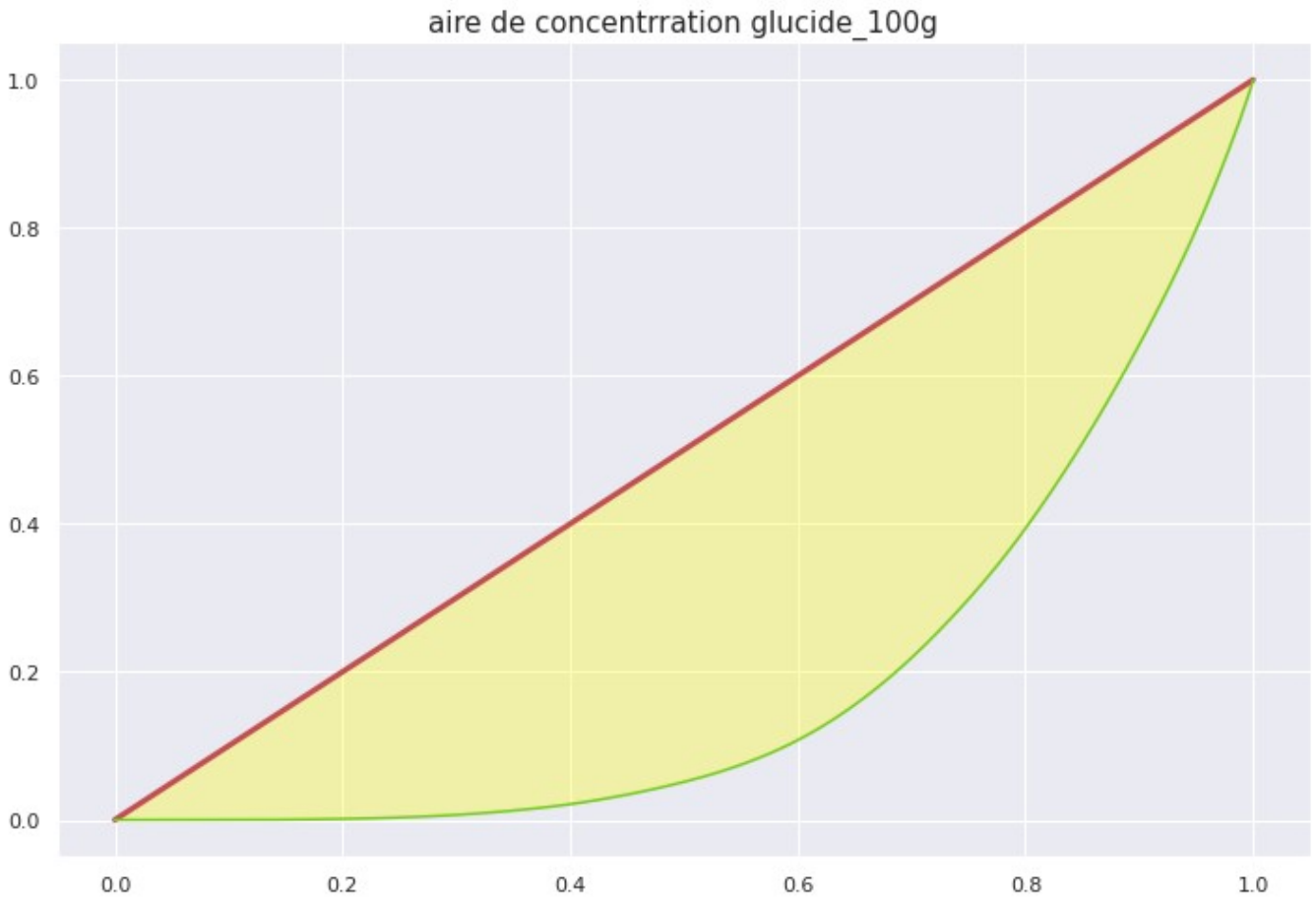




Repartions des Gluciscore

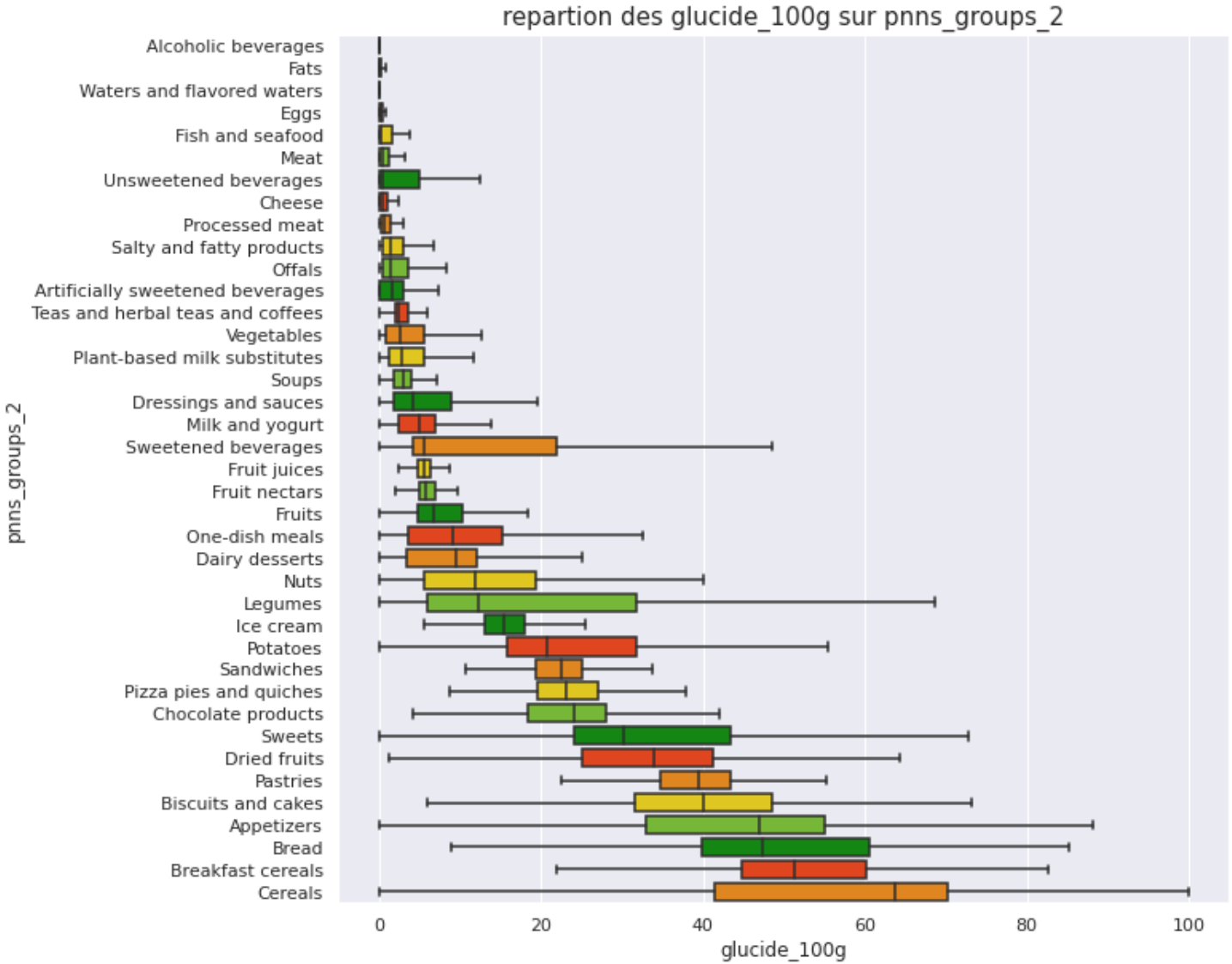


Repartions des glucides par aire de Gini



aire de gini 0.6190368314738456

Repartions des glucides par pnn group 2



Information très importante

car il va nous permettre de développer un algorithme de recommandations basé sur ces informations!!

Repartions des Gluciscore sur le groupe pnn group 1

pnns_groups_1 glusiscore_grade	Alcoholic beverages	Beverages	Cereals and potatoes	Composite foods	Fat and sauces	Fish Meat Eggs	Fruits and vegetables	Milk and dairy products	Salty snacks	Sugary snacks
a	3138	29	5270	116	13	1299	5610	8097	2359	4252
b	79	1720	9196	1062	969	2078	7880	17757	8527	23303
c	158	1113	4907	5045	3151	12406	1803	777	2274	8773
d	418	12808	0	11527	7205	17294	0	215	348	7703
e	0	2688	0	3866	1635	0	0	0	0	688

OpenFoodFacts

Le test de Fisher nous à montrer que nos variables qualitatives(pnns_group1 et 2) ont une influence sur nos variables quantitatives

Le test de normalité Kolmogorov Smirnov nous révèle que les distributions ne sont pas normales

Le test ANOVA(kruskal) sur les variantes non paramétriques

nous indique que nos distributions ne sont pas les mêmes

Le test de Friedman nous indique que nos distributions ne sont pas les mêmes

Les tests Chi 2 ainsi que le tableau corrélation linéaire nous ont révélé des dépendances entre les variables

Le méthode k-means ne nous a pas aider à regrouper nos données

Source OpenFoodFacts

Questions/Réponses

Source OpenFoodFacts

Thank you!