Formation DATA SCIENTIST PROJET 2

Analysez des données nutritionnelles
Soutenance du 16/06/2021

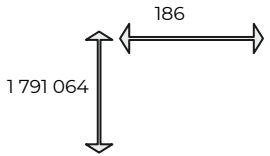
Sommaire

Problématique

Le site Lamarmite souhaite construire un générateur de recettes saines. Je dois réaliser **une analyse de données exploratoire** d'une base de données mettant à disposition des ingrédients avec leurs caractéristiques nutritionnels.

Cette base de données comporte 1 791 064 ingrédients avec 186 caractéristiques

Il faut réaliser un **nettoyage** puis une **exploration** afin d'en déduire les ingrédients "sains" à utiliser dans des recettes



Description des données

Il y quatre grandes catégories de données

- Les informations générales sur la fiche du produit :
  a. nom
  b. date de modification
- Un ensemble d'étiquettes
  a. catégorie du produit
  b. localisation, origine
- Les ingrédients composant les produits et leurs additifs éventuels
- **Des informations nutritionnelles** : quantité en grammes d'un nutriment pour 100 grammes du produit

Toutes ces informations ont un taux de valorisation très différent. Un premier travail a consisté à repérer ces taux afin de supprimer les caractéristiques non renseignées pour être utilisées.

Nettoyage

Constats :

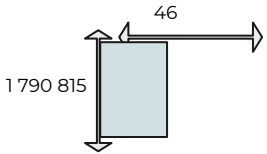
- Une grande majorité d'attributs très peu valorisé ⇒ Mise en évidence à l'aide du graphique suivant

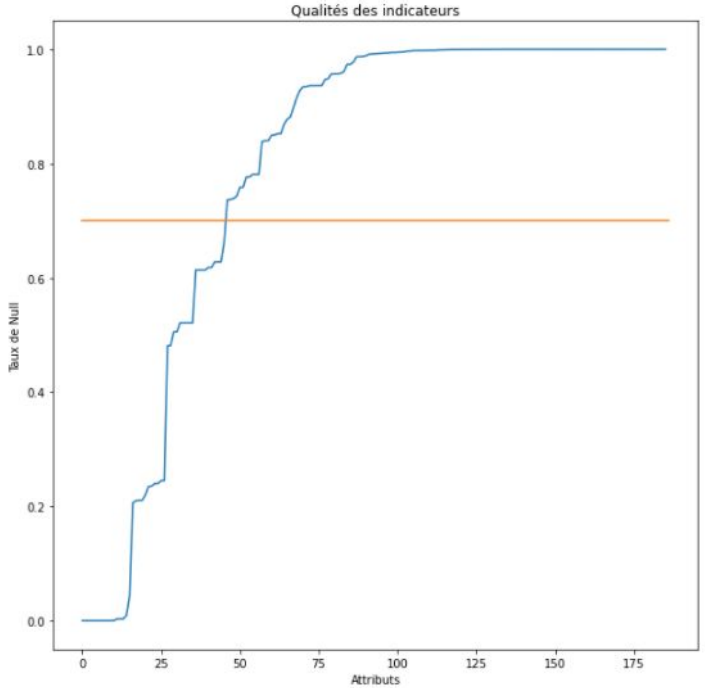
⇒ Ne pas retenir les attributs avec un taux de nullité supérieur à 70%

- Un attribut permet d'identifier de façon unique l'ingrédient

⇒ Une suppression des doublons est réalisée

Notre jeu de données subit une réduction importante des caractéristiques

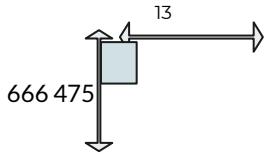




Nettoyage

Conservation	***Suppression***
Code produit afin d'identifier de façon unique notre ingrédient	Informations sur le **cycle de vie du produit** (Date de création, créateur etc…) qui ne permettent pas d'apporter des éléments sur la qualité nutritionnelle du produit
Information permettant de classer la **catégorie** du produit (Viande, Poisson etc…)	Lignes **sans score nutritionnel** a été réalisée. Ce point a fortement réduit le jeu de données (63%) mais il paraissait difficile de créer une règle afin de le déduire à ce stade. Les résultats ne semblent pas avoir été faussés par cette opération.
Informations nutritionnelles (Sucre, Graisse etc.. pour 100gr)	
Score nutritionnel : Score ainsi que le Grade	

```
Index(['code', 'pnns_groups_2', 'energy_100g', 'proteins_100g',
       'saturated-fat_100g', 'energy-kcal_100g', 'salt_100g', 'sodium_100g',
       'carbohydrates_100g', 'sugars_100g', 'fat_100g', 'nutriscore_grade',
       'nutrition-score-fr_100g'],
```



Nettoyage

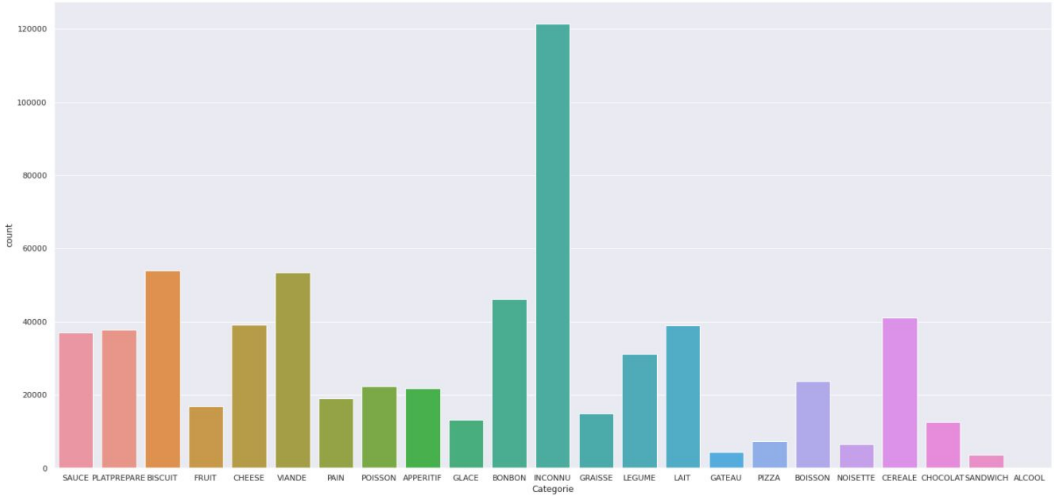
Notre jeu de données présente toujours deux éléments qu'il faut corriger pour la suite de l'analyse :

- Des valeurs nulles
- Des valeurs aberrantes

⇒ Ces éléments sont présentés ici afin de suivre leur correction

```
code                            0
pnns_groups_2                   0
energy_100g                  1673
proteins_100g                1715
saturated-fat_100g           1735
energy-kcal_100g            41603
salt_100g                    1305
sodium_100g                  1305
carbohydrates_100g           1988
sugars_100g                  1727
fat_100g                     1716
nutriscore_grade                0
nutrition-score-fr_100g         0
dtype: int64
Valeurs Aberrantes fat_100g  :  14
Valeurs Aberrantes salt_100g  :  224
Valeurs Aberrantes sugars_100g  :  27
Valeurs Aberrantes proteins_100g  :  14
Valeurs Aberrantes sodium_100g  :  142
Valeurs Aberrantes saturated-fat_100g  :  7
Valeurs Aberrantes carbohydrates_100g  :  59
```

Nettoyage



Valeurs nulles

Afin d'imputer les valeurs nulles nous avons procédé au traitement suivant :

- Détermination d'une catégorie de produit : 23 Catégories ont été déterminées ⇒ Expliquer comment ?
- Imputation de la valeur manquante par la médiane de la catégorie du produit

```
code                        0
pnns_groups_2               0
energy_100g                 0
proteins_100g               0
saturated-fat_100g          0
energy-kcal_100g            0
salt_100g                   0
sodium_100g                 0
carbohydrates_100g          0
sugars_100g                 0
fat_100g                    0
nutriscore_grade            0
nutrition-score-fr_100g     0
Categorie                   0
```

Nettoyage

Les valeurs aberrantes n'ont pas été traitées avec la méthode de l'IQR car elles ne suivaient pas une répartition Gaussienne

Valeurs aberrantes

Pour les valeurs aberrantes il n'a pas été décidé d'imputer mais de supprimer les enregistrements concernés car leur volume était très faible.

Afin de les détecter des règles métiers ont été définies :

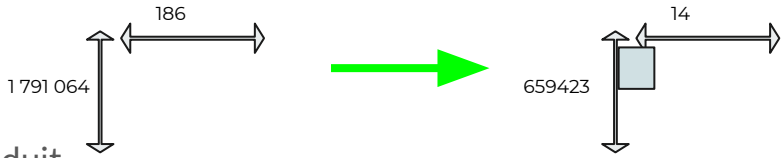
- La plupart des indicateurs indiquent un taux pour 100gr de produit, par conséquent tous les produits présentants des taux strictement inférieur à 0 ou supérieur à 100 ont été supprimés
- Deux indicateurs sont différents : Energie pour 100g et l'énergie en KCAL. Pour ces deux indicateurs des recherches sur les aliments les plus énergétiques ont permis d'identifier les taux à retenir.
    - Energie pour 100g ⇒ max 3600
    - Energie pour 100g en KCal ⇒ max 850

```
Valeurs Aberrantes fat_100g  :  0
Valeurs Aberrantes salt_100g  :  0
Valeurs Aberrantes sugars_100g  :  0
Valeurs Aberrantes proteins_100g  :  0
Valeurs Aberrantes sodium_100g  :  0
Valeurs Aberrantes saturated-fat_100g  :  0
Valeurs Aberrantes carbohydrates_100g  :  0
Valeurs Aberrantes energy-kcal_100g  :  0
Valeurs Aberrantes energy_100g  :  0
```

Nettoyage

Synthèse

- Notre jeu données est totalement nettoyé.
- Aucune valeur nulle ni de valeur aberrante.
- Un attribut permet d'identifier la catégorie de produit
- Quatorze dimensions retenues
- Environ 650 000 enregistrements



⇒ Ecriture d'un fichier de sortie pour l'utiliser dans le traitement d'Analyse

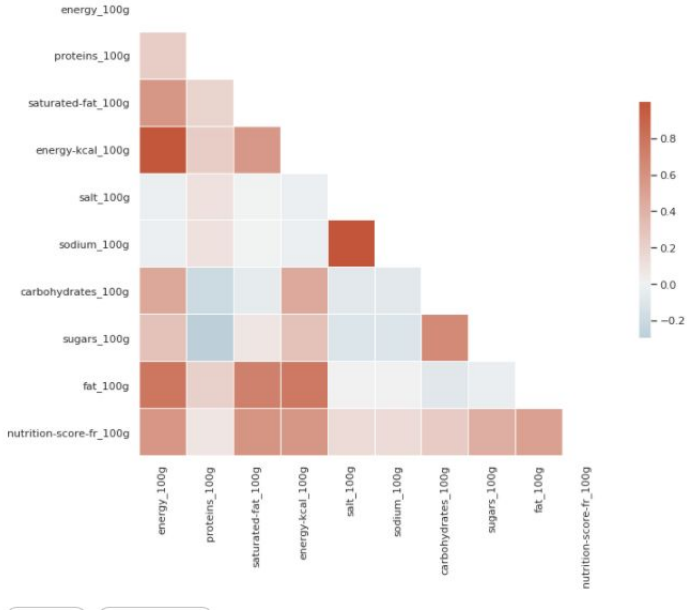
Analyse

Une première approche a été de calculer la matrice de corrélation entre les variables afin de supprimer celles parfaitement corrélées.

Constat : Quatre variables parfaitement corrélées deux à deux :

- Salt 100g avec le Sodium 100g
- Energy 100g avec Energy Kcal 100g

⇒ Nous allons supprimer le Sodium 100g et l'Energy Kcal 100g pour la suite des traitements

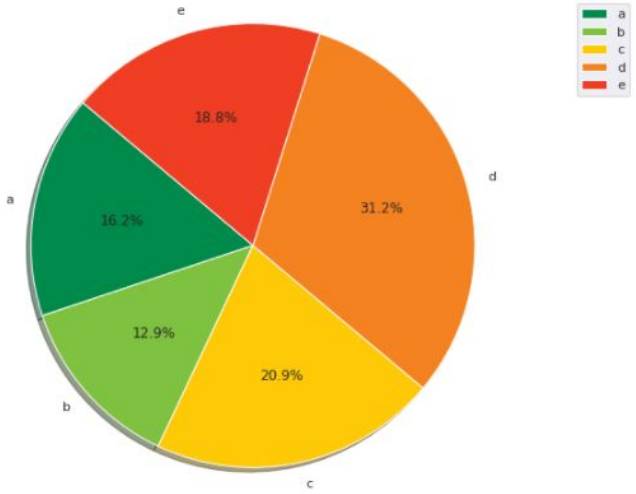


Analyse

Réalisation d'un **Pie Chart** pour présenter la répartition des produits par nutri score

Constat : La majorité des produits ont un nutri score entre C et E

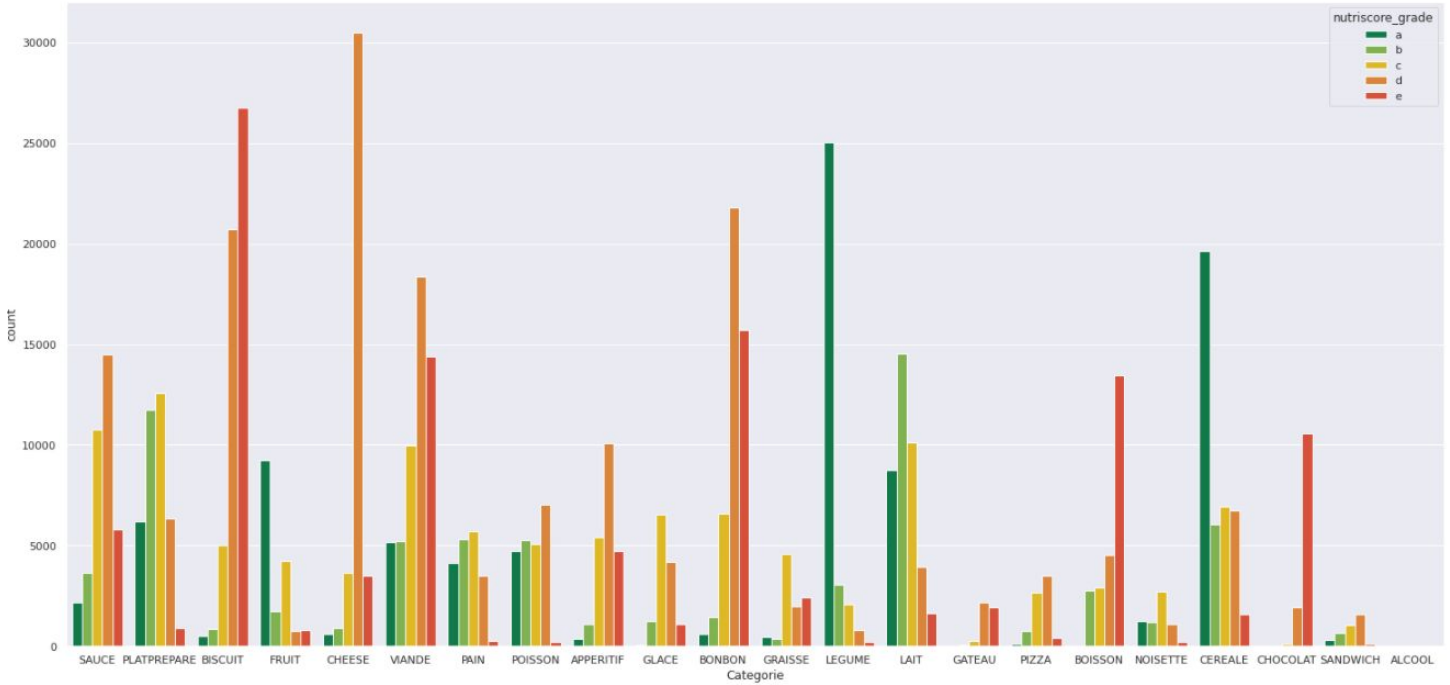
Nous pouvons les représenter différemment grâce à notre catégorie créée.



Analyse

Réalisation d'un **Barplot** afin d'identifier le nombre de produits par Catégorie et par Nutri Grade.

Constat : Nous pouvons très vite constater la différence entre les catégories.



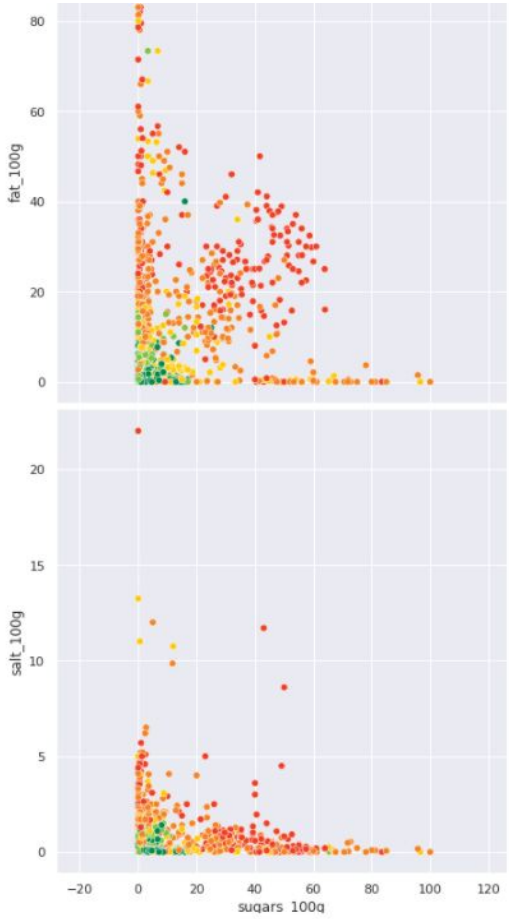
Analyse

Réalisation d'un **Pair Plot** entre le sucre, la graisse, le sel suivant le nutri score.

Le Pair Plot nous permet de voir les corrélation deux à deux entre des dimensions en ajoutant une troisième avec le nutri score.

Constat : le Pair Plot entre le sucre et la graisse est très représentatif pour l'évaluation du nutri score. Plus il y a de sucre et de graisse et plus le score sera mauvais

Nous constatons la même chose entre le sucre et le sel mais dans des proportions moins importantes

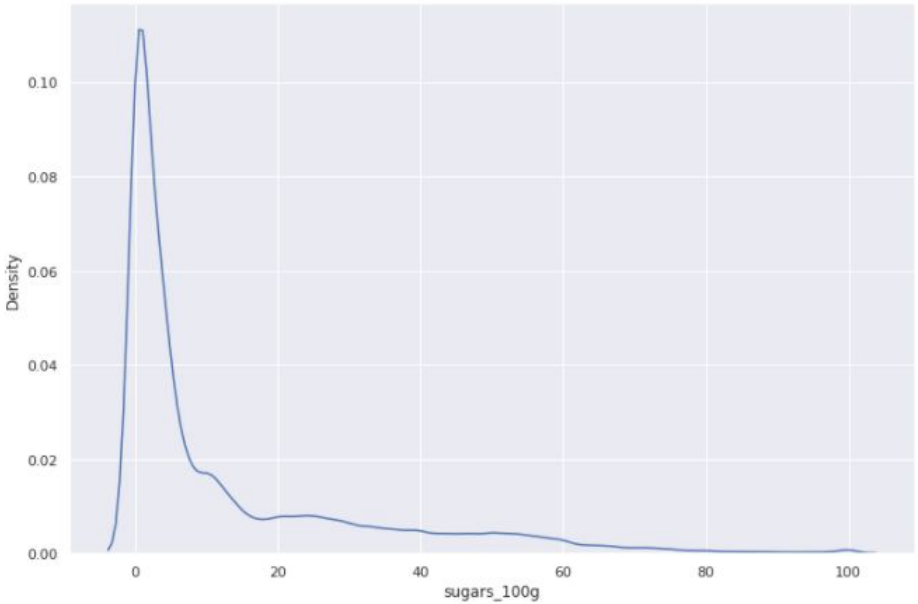


Analyse

Réalisation d'un **KDEPlot** sur le Sucre.

Constat : la répartition ne suit pas une courbe de Gauss ou en cloche.

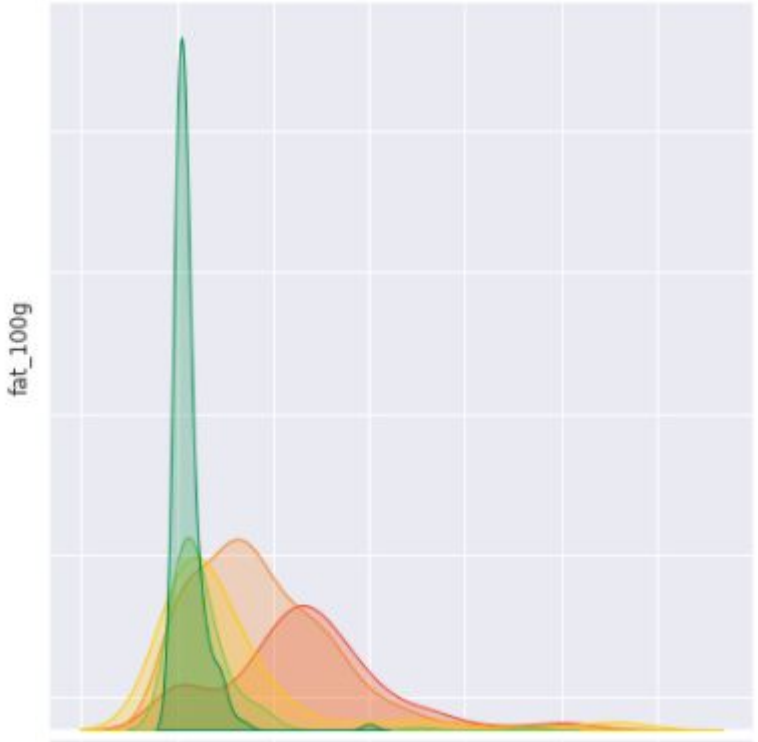
Cette répartition cible davantage le choix de méthodes pour trouver des corrélations fortes entre les variables.



Analyse

Le Pair Plot a généré des KDE Plot sur les trois dimensions demandées Sucre / Graisse / Sel.
En ajoutant le Nutri score comme dimension supplémentaire

Constat : Celui de la graisse est le plus parlant afin d'identifier son influence dans le nutri score



Analyse

Nous pouvons regarder plus précisément le sucre afin de visualiser son importance dans le score nutritionnel.

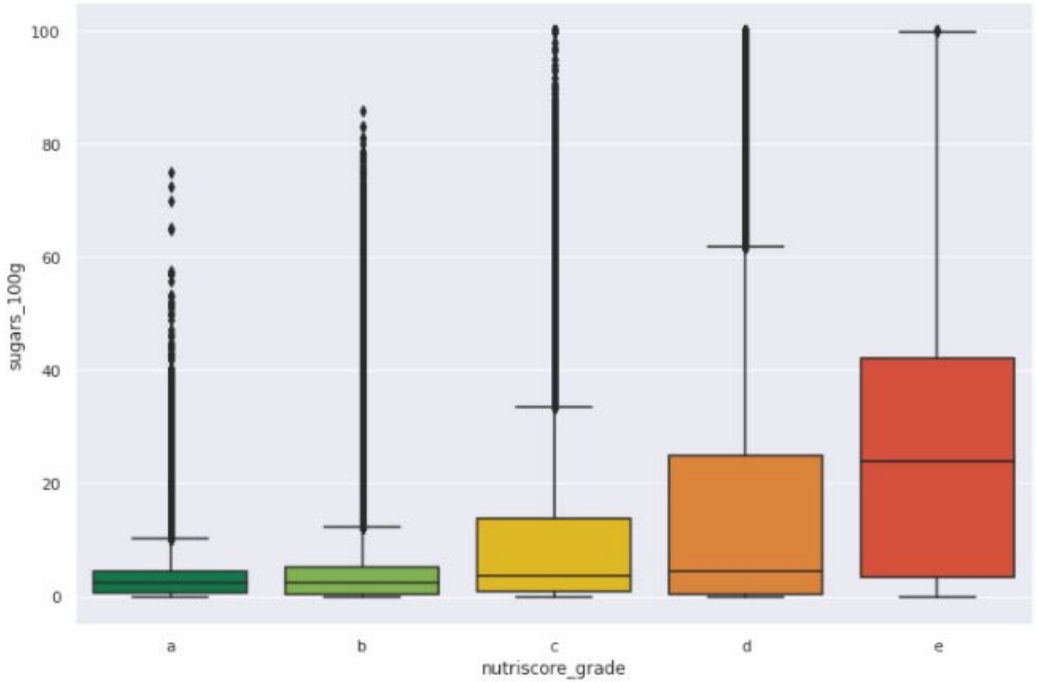
Pour cela nous avons réalisé un **Boxplot** par nutri score.

Constat : l'importance du sucre avec des médianes qui grimpent plus le score se dégrade

Ce point est confirmé par le **test de Kruskal-Wallis**

Le test de Kruskal vérifie H0 dans le cas ou les médianes de nos échantillons sont identiques. Sinon celui vérifie H1.

Le P-Value de 0 obtenu confirme l'hypothèse H1 donc nos échantillons se basant sur le nutri score sont bien différents



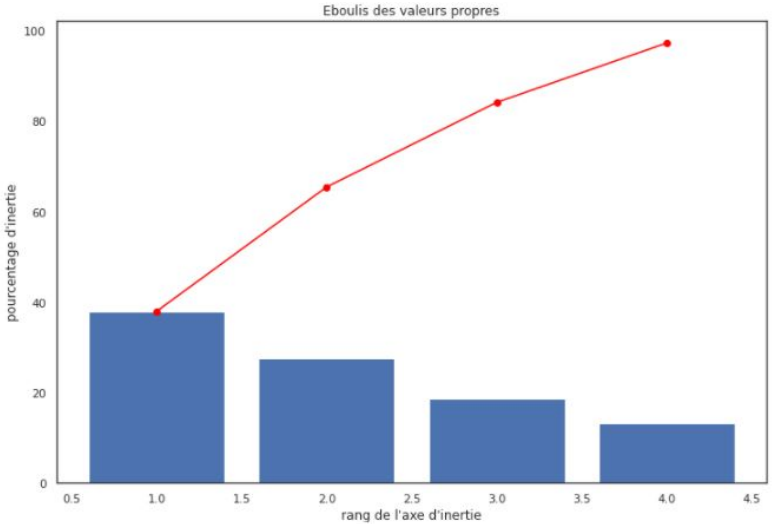
Analyse

Réalisation d'une Analyse en Composante principale (ACP)

Hypothèse

- Nous allons conserver uniquement les cinq dimensions quantitatives
- Calcule de quatre composantes principales F1, F2, F3 et F4

⇒ Sur l'éboulis suivant nous pouvons constater une couverture de presque 100% sur nos quatre composantes principales



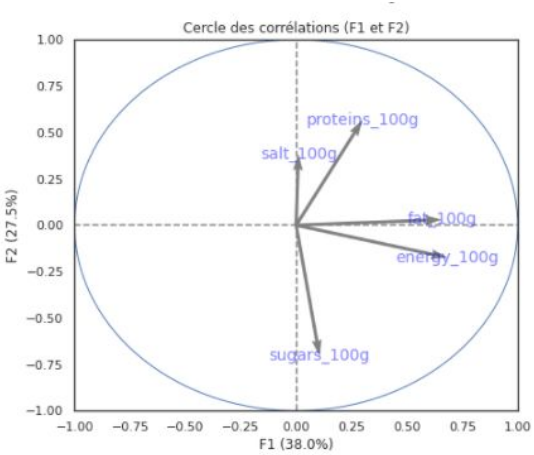
Analyse

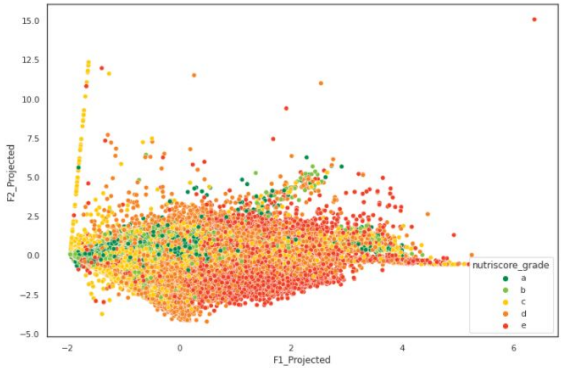
Cercle des corrélations de F1 et F2

- Sur l’axe F1 la teneur en graisse et l'énergie dans 100g jouent un rôle déterminant
- Sur l’axe F2 c’est la teneur en sel et en sucre qui sont déterminants mais de façon opposée

Interprétation avec la projection des individus

- Nous pouvons voir grâce à la projection des individus une répartition par nutri grade
- Les individus sur la gauche ont un taux de graisse et d'énergie important
- Ceux qui s’écartent vers le bas ont un taux de sucre plus important
- Ceux qui s’écartent vers le haut ont un taux de sel plus important





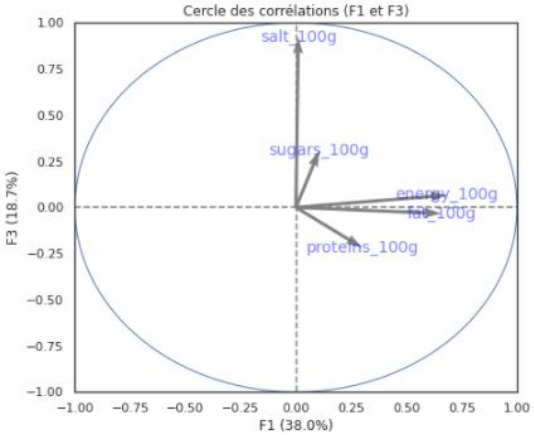
Analyse

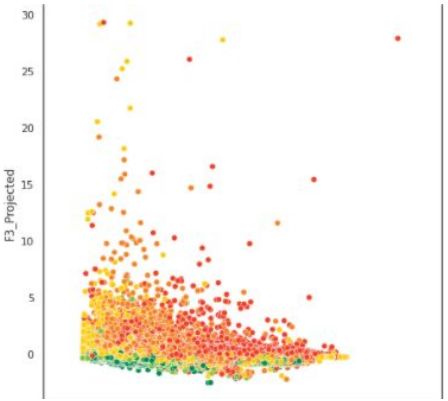
Cercle des corrélations de F1 et F3

- Sur l'axe F1 la teneur en graisse et l'énergie dans 100g jouent un rôle déterminant
- Sur l'axe F3 c'est la teneur en sel et en sucre qui sont déterminants et vont dans le même sens

Interprétation avec la projection des individus

- Nous pouvons voir grâce à la projection des individus une répartition par nutri grade.
- Les individus sur la gauche ont un taux de graisse et d'énergie important
- Ceux qui s'écartent vers le haut ont un taux de sel et de sucre plus important





Conclusion

Grâce à l'ACP nous avons pu déterminer des composantes principales qui nous permettaient de "classifier" clairement les individus par nutri grade. Cela devrait permettre une prédiction facilité de celui ci lorsque l'information est manquante.
Sans surprise les informations nutritionnelles permettant de le définir sont le **Sucre, le Sel, les Graisses et l'énergie**.

Par conséquent il est simple de faire un traitement qui prendra ces éléments pour prédire le nutri grade et ainsi viser des produits uniquement en catégorie A ou B.