

Parcours DataScientist Projet 2:

Analyse de données Exploratoire

pour le site Lamarmite



Jérôme d'Harveng

Mentor: Pierre Comalada

OpenClassrooms janvier 2019

Table des matières

PART I: Contexte de l'analyse de données

PART II: Nettoyage de données

PART III : Analyse exploratoire

PART IV: Conclusions, différentes pistes



Part I:

Contexte de l'analyse



Problématique



- *Lamarmite* :

* But final: un générateur de recettes saines

* Informations:

une BD

avec nombreux produits de consommation

* Points importants:

Avantages / Inconvénients nutritionnels des aliments



Interprétation

- Brève recherche sur les bases d'**une alimentation saine**

* vue comme **favorable**:

légumes/fruits, produits pêche, vitamines, minéraux

* vue comme **moins favorable**:

Produits gras, sucrés ou salés



- Recherche description champs BD

<http://fr.openfoodfacts.org/score-nutritionnel-experimental-france>

* Importance « **Score nutritionnel** » (Food Standards Agency)

* Travail professeur Hercberg

Quelques pistes de recherche

- Se focaliser dans un premier temps sur :

- * le score nutritionnel
- * le niveau d'énergie (kJ)
- * les caractéristiques principales
tq lipides, glucides, fibres, protéines, sel
- * les vitamines et minéraux



Part II:

Nettoyage des données



Découverte de la Base de données

- **BD** de (320772, 162) => **trop** de colonnes
 - * une première sélection d'après les pistes (PART I)
 - * parmis les 4 catégories => informations nutritionnelles

Plupart des données en [g/100g] ou [kJ/100g]

- Utilisation de la librairie **Pandas** de Python pr:
 - * avoir un aperçu des données : head(), tail(), sample()
 - * voir le type des variables : info(), dtypes()
 - * rechercher les valeurs manquantes: isnull()
 - * rechercher les valeurs aberrantes: describe()

Premières observations

- Besoin d'uniformiser les noms des colonnes (- vs _)
- Pas de doublons

- **Valeurs erronées:**

Valeurs <0 ou > 100g pour les nutriments _100g

Valeurs > 1g pour les vitamines et minéraux

- **Valeurs manquantes:**

- * code (23 rows)
- * product_name (17762 rows, 5.5%)
- * generic_name (267977 rows, 83.5%)
- * nutriments ending with _100g
- * nutrition_score_fr_100g et nutrition_score_uk_100g
- * nutri_grade_fr_100g

- **Types:**

- * nutrition_score_fr_100g (and...uk...) est en *floats*, mais devrait être *int*.
- * nutrition_grade_fr est *String*, mais devrait être une variable *discrète ordonnée*

Valeurs manquantes

Il existe **différentes solutions** pour le nettoyage



- **Supprimer les lignes des produits associés**

Si le nombre de valeurs manquantes pas trop important comme pour : « code » et « product_name »

- **Supprimer la variable concernée (colonne)**:

Si le nombre de valeurs manquantes est trop grand et variable pas indispensable pour l'étude comme pour « generic_name »

- **Remplacer les valeurs manquantes**:

Pour « ... _100g » : hypothèse si NAN => non présent => 0

Pour « nutrition_score_X_100g » => par la médiane (moins sensibles aux valeurs abérantes)

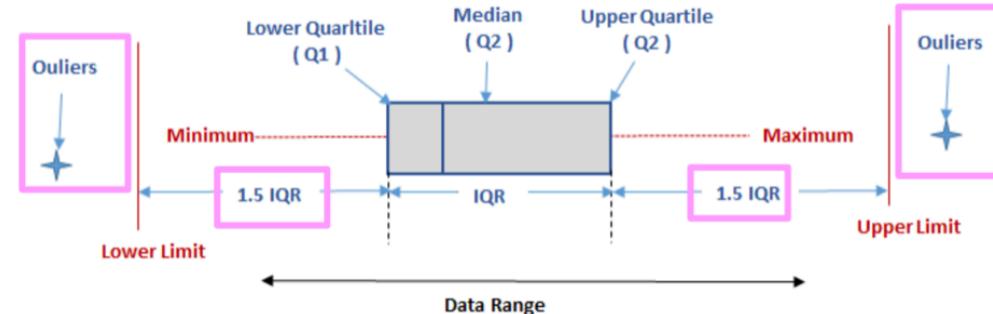
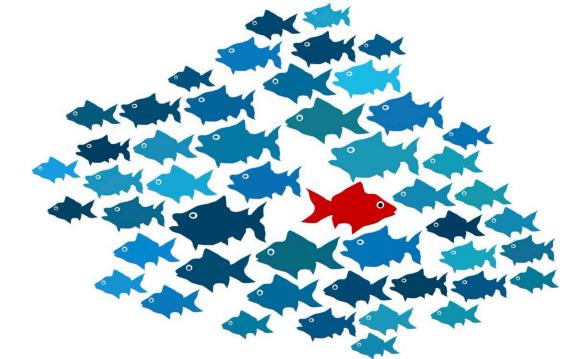
- **Remplacer et marquer les valeurs remplacées**:

Afin de pouvoir filtrer les valeurs remplacées lors de l'exploration de « nutrition_grade_fr » => ajout d'une colonne additionnelle avec des 0 et 1

Valeurs aberrantes

- Détection

Via les BoxPlots



Remarque :

on s'est concentré surtout sur « nutrition_score_X_100g »
et « energy_100g ». Car cette méthode peut donner mauvaises indications si
beaucoup de « 0 ».

- Traîtement

Si étude comportement particulier => on peut les garder (ex. Détection de fraude)
Ici pas le cas => **Supprimer** les produits concernés (lignes)

Part III:

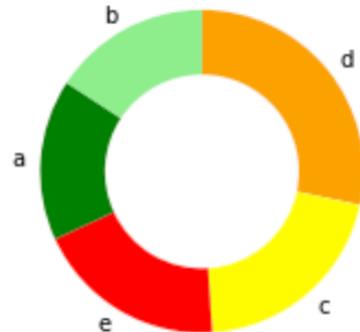
Analyse Exploratoire

Analyse Univariée

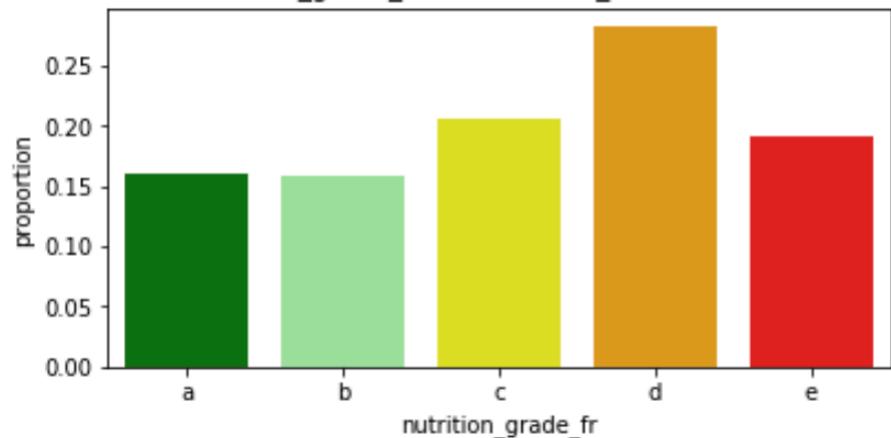


nutrition_grade_100g

Repartition of nutrition grades according to amount of aliments



"nutrition_grade_fr" with all null_values removed

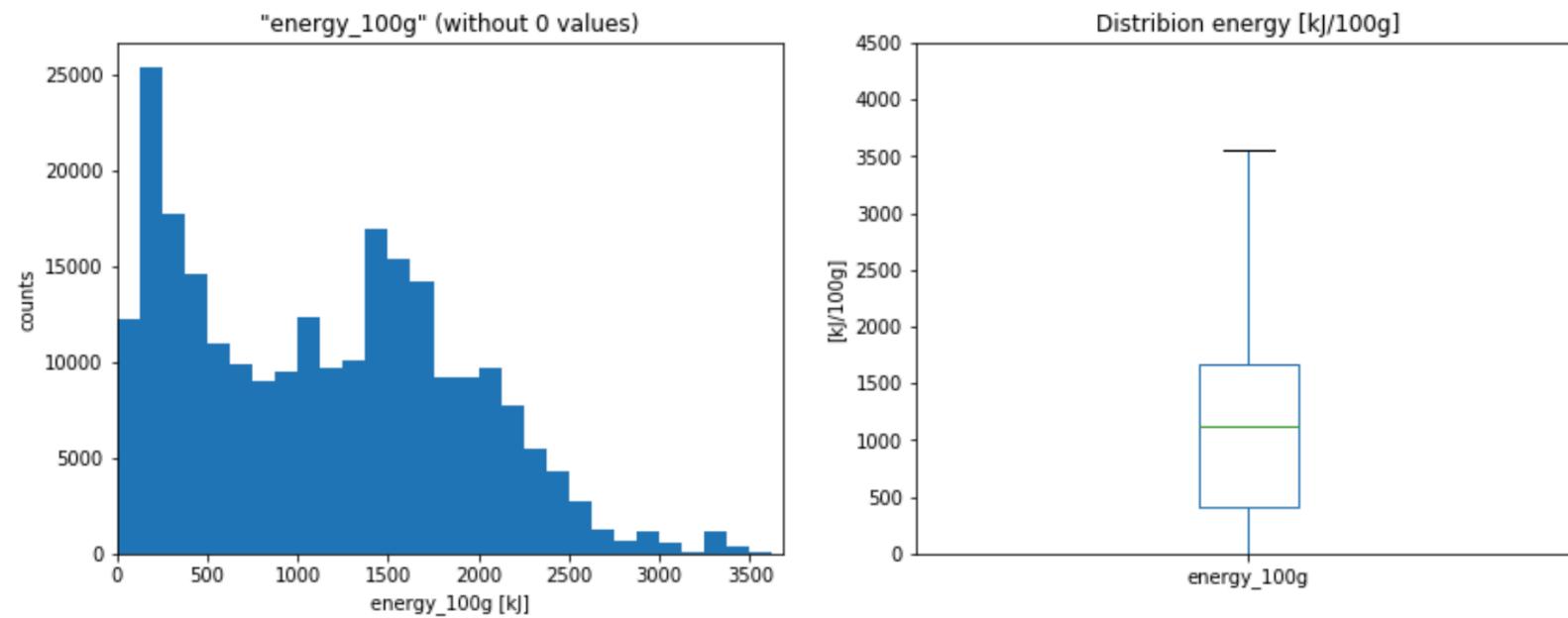


Observations:

- * Catégorie « d » : contient le plus d'aliments
- * Sinon le reste des aliments ~ uniformément réparti

Analyse Univariée

energy 100g



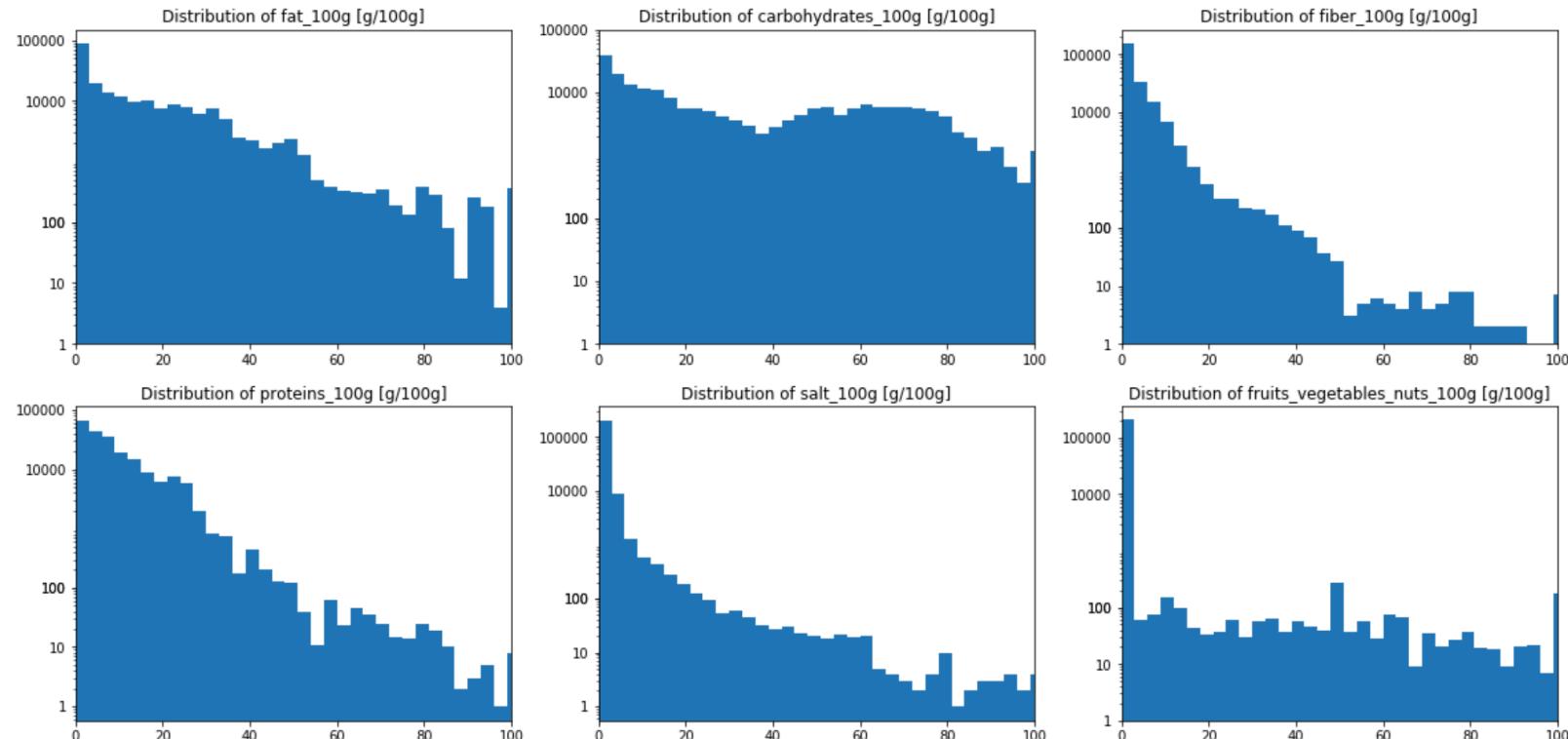
Observations:

- * distribution « energy_100g » ~ bimodal (125-250kJ et 1500-1625kJ)
- * 50% < 1120 kJ et 75% < 1674kJ

Analyse Univariée

Axes des y en échelle logarithmique

Main nutriments 100g: fat, carbohydrates, fiber, proteins, salt and fr_vgt_nuts



Observations:

- * Distributions pour graisses, fibres, protéines et sel semblent « Right-Skewed »
- * Distributions pour les carbohydrates et fruits et légumes plus uniformes.

Feature Selection

Retirer les « features » intercorrélées:

Matrice de corrélation (critère > 0.9)

=> retrait «sodium_100g » et « nutrition_score_uk_100g »



Sélection des « features » importantes:

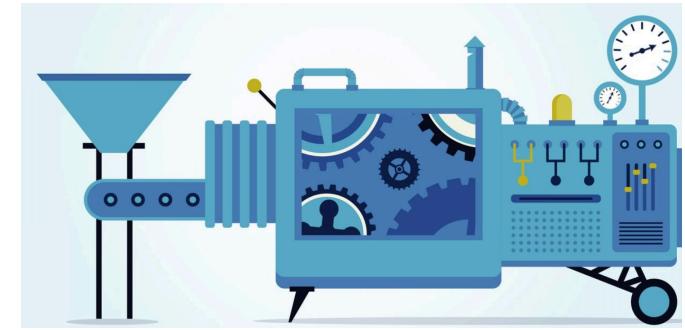
Random Forest (\mathbf{X} = paramètres déductions, \mathbf{y} = « nutrition_score_fr_100g »)

Permet d'analyser comment se comporte le modèle en retirant une variable

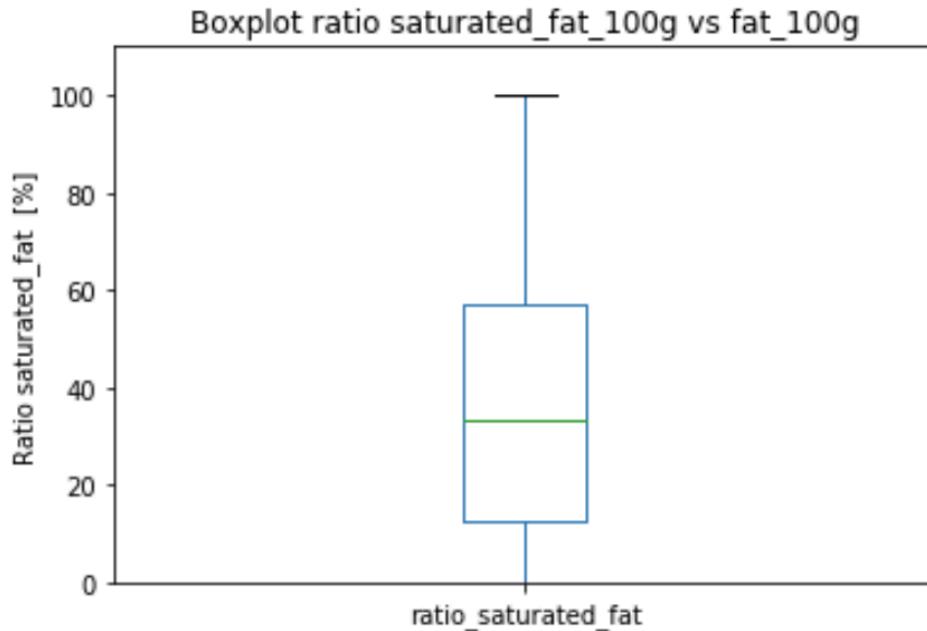
=> Garder:

- salt_100g, carbohydrates_100g, proteins_100g, fat_100g,
saturated_fat_100g, fiber_100g
- iron_100g, calcium_100g, potassium_100g
- vitamin_a_100g, vitamin_c_100g

Feature Engineering



Ratio : saturated_fat_100g / fat_100g



Permet de différencier type de graisses (huiles, margarines, crèmes fraîches) seuil > 10 pr ratio

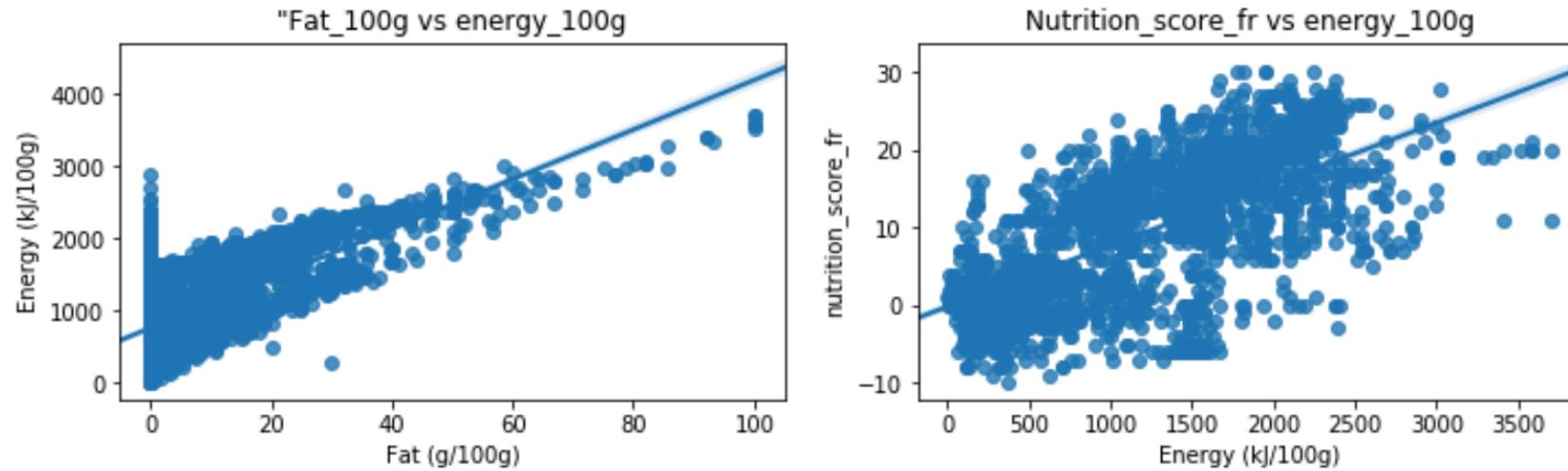
Observations:

25% aliments < 12,49%

=> moins de 25% aliments < 10%

Analyse Multivariée

energy 100g, fat 100g and nutrition score fr



Corrélation linéaire :

Pearson coefficient : energy vs fat

0.7233825237946157

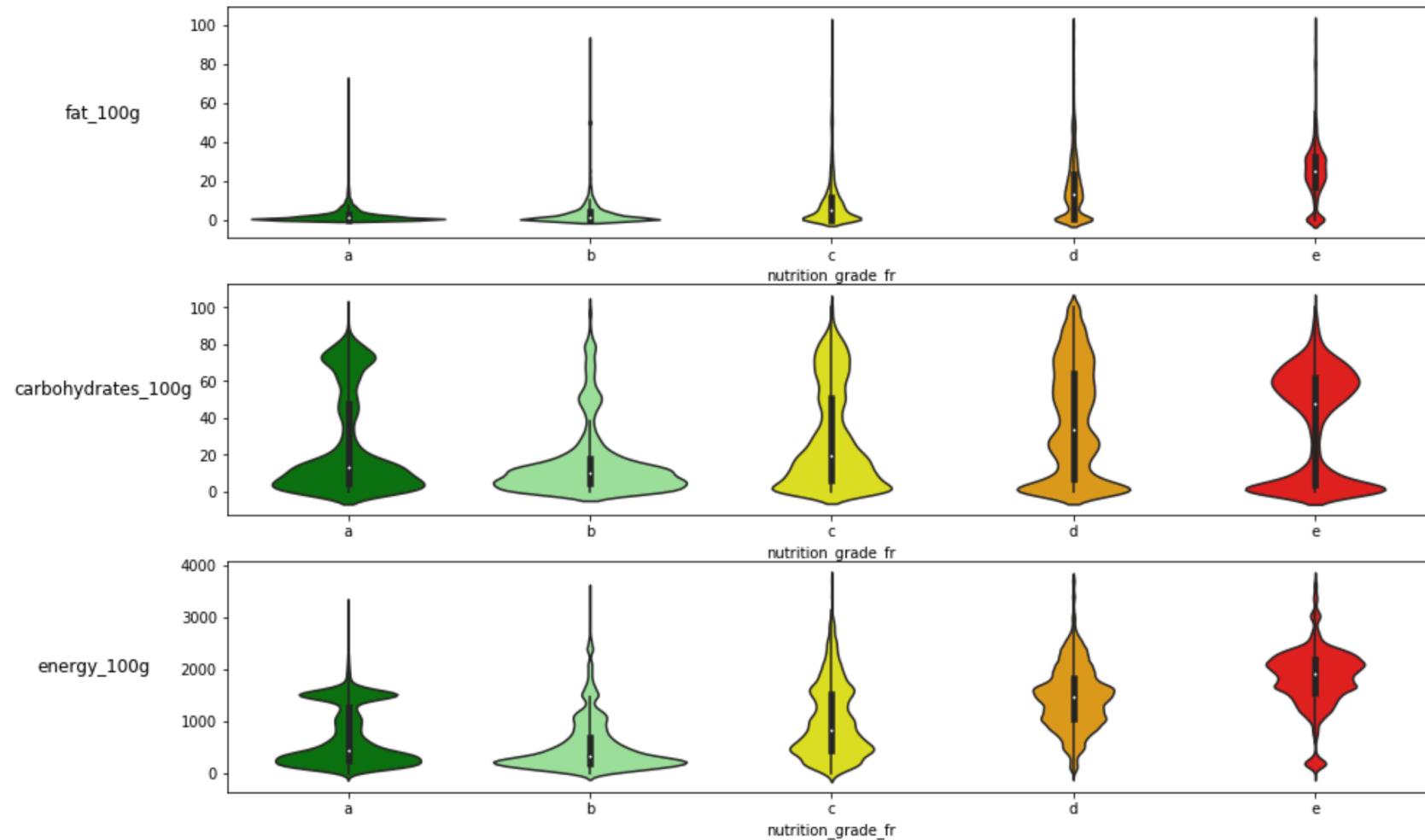
Pearson coefficient: energy vs nutrition_score_fr

0.6381268973153934



Analyse Multivariée

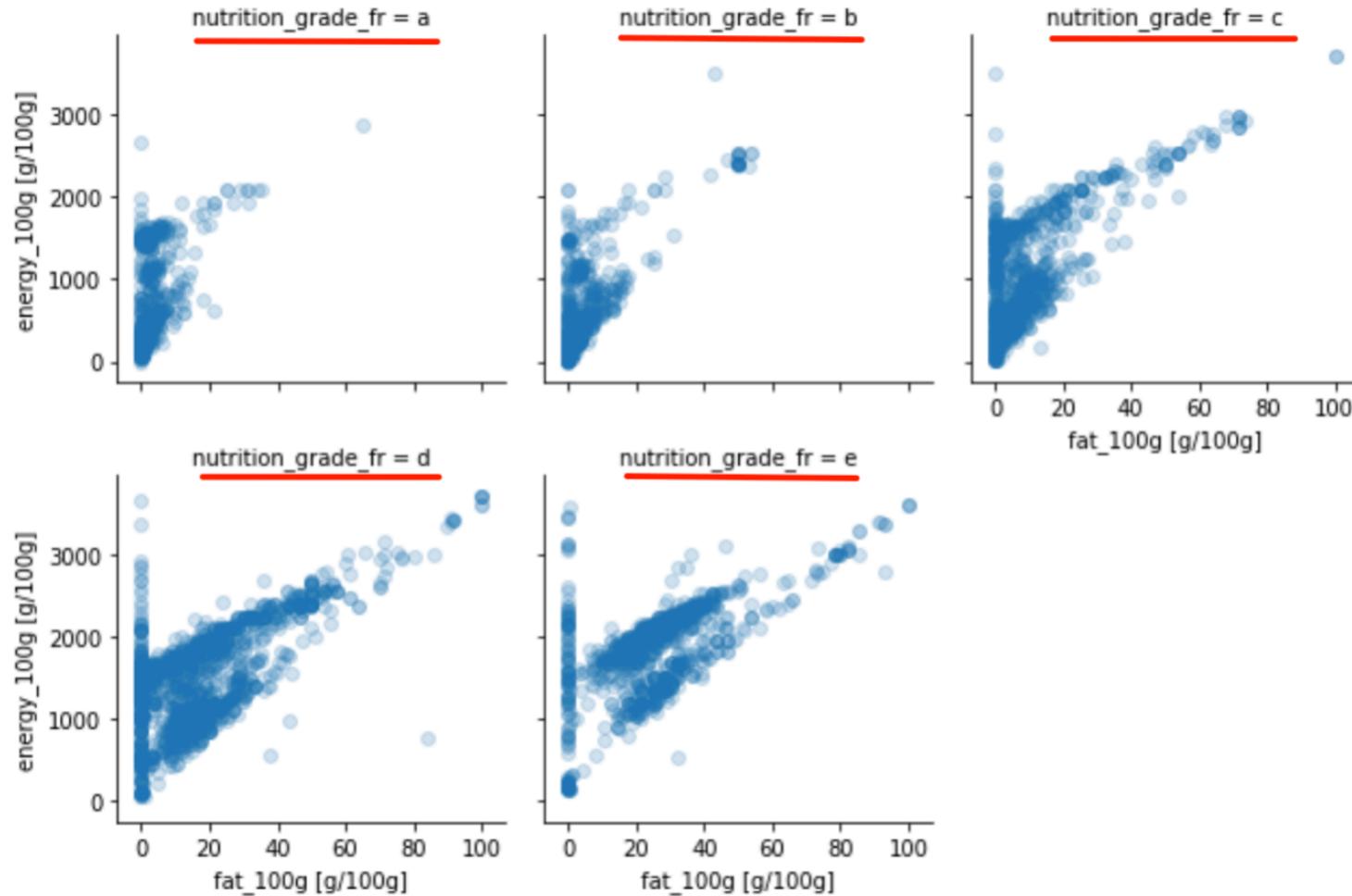
Violinplots => distribution: fat, carbohydrates et energy





Analyse Multivariée

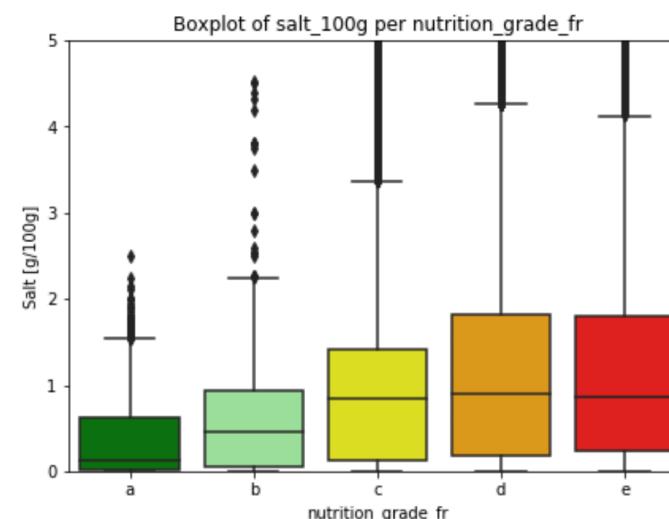
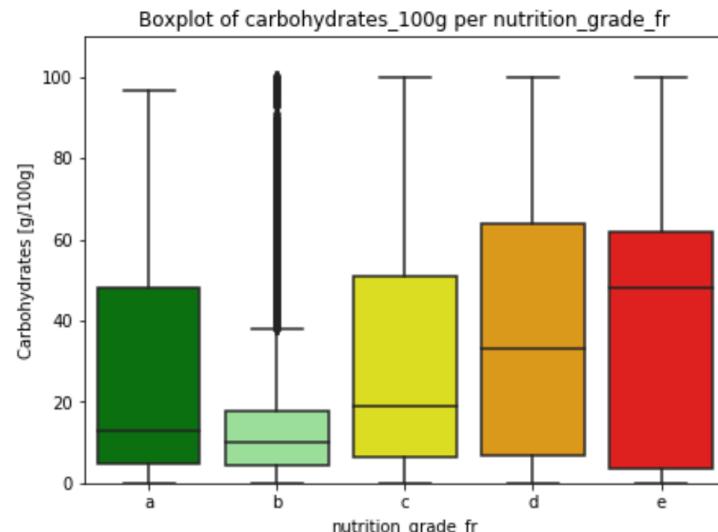
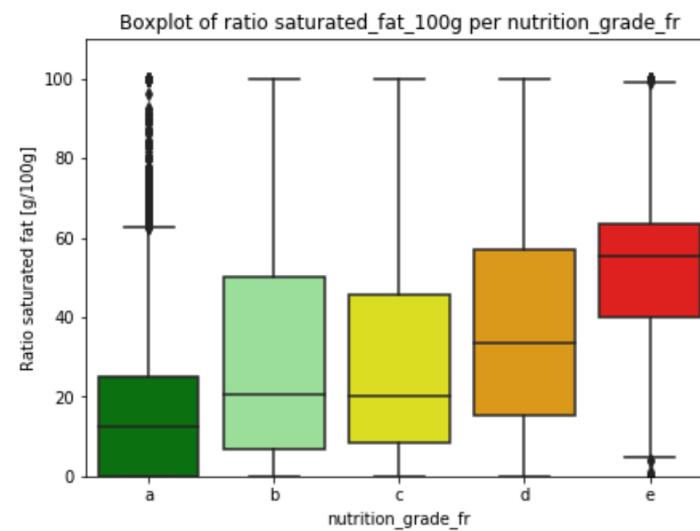
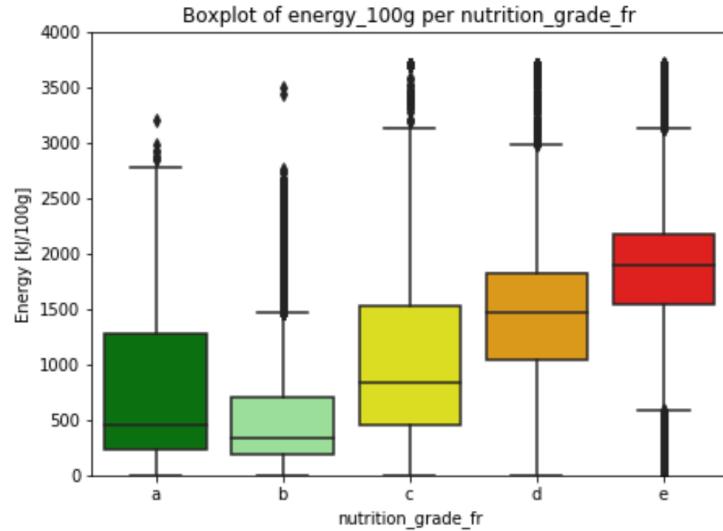
Fat vs energy 100g pour chaque grade



Analyse Multivariée



Boxplots en fonction de la note nutritionnelle: Energy / ratio saturated_fat / carbohydrates / salt

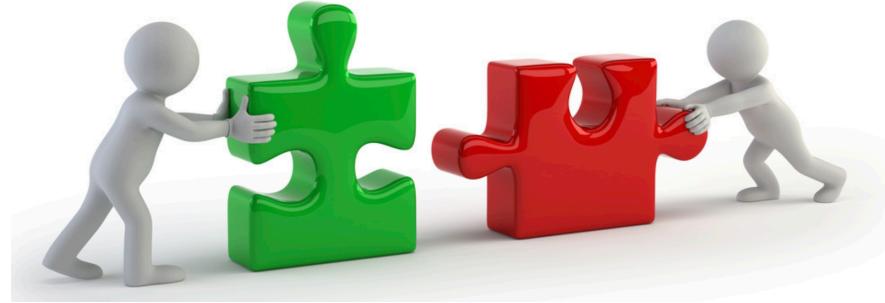


Part IV:

Conclusions

Conclusions

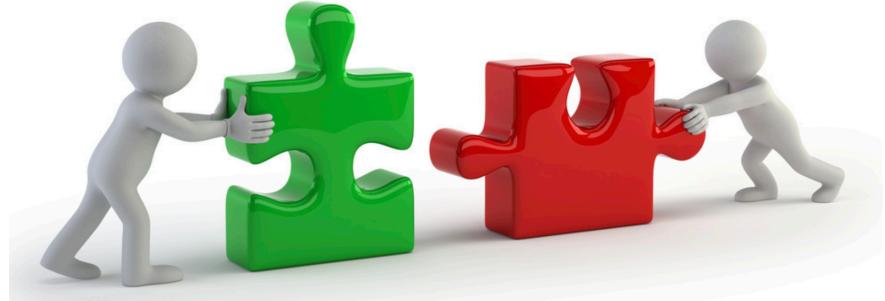
Pour rappel:



On s'est basé sur l'étude du *professeur Hercberg*

- * importance du *score nutritionnel* et du *grade associé*
- * *grades de « a » (sain) à « e » (moins sain)*
- Afin de faciliter la tâche de Lamarmite
 - => Diminution du nombre de variables d'intérêt
- Avt d'utiliser la BD, un nettoyage des données a été réalisé:
 - => **détection/traitemen**t valeurs erronées, manquantes, aberrantes
- A l'aide de l'**étude exploratoire**, on a pu observer que:
 - * le grade contenant le plus d'aliments = « d »
 - * à l'aide du Feature Selection, on a décidé de se focaliser sur le niveau d'énergie, *les graisses, les carbohydrates et le sel*

Conclusions



- * De plus, nous avons ajouté un ratio (entre graisses saturées et graisses totales) => pr différencier les graisses en provenance des huiles, margarines ou crèmes fraîches
- * Les aliments plus sains (grades « a » et « b ») ont :
 - # un niveau d'énergie [kJ/100g] plus faible
Médiane < 448 kJ/100g
 - # un ratio de graisses saturées plus faible
Médiane < 20% (en comptant « c »)
 - # un niveau de carbohydrates plus faible
Médiane < 13g/100g
 - # un niveau de sel plus faible
Médiane < 0.46 g/100g

Pistes

Lors de l'élaboration de recettes saines:

Lamarmite pourrait **se concentrer** sur des *ingrédients* avec un **niveau énergétique inférieure**, et une **bonne combinaison** des niveaux **de graisses, carbohydrates et sel.**

D'autres nutriments tels que les **protéines** ne sont pas spécifiques à un grade en particulier.



