OPENCLASSROOMS

CentraleSupélec

Concevez une application au service de la santé publique

Formation Data Scientist - Janvier / Novembre 2022

Auteur: Eric TREGOAT

Mentor: Benjamin TARDY

Evaluatrice: Fatou SALL

Ordre du jour de la soutenance

- **Présentation (20 minutes)**
  - Rappel de l'appel à projets et explication de votre idée d'application (2 mn)
  - Démarche méthodologique de nettoyage (8 mn)
  - Démarche méthodologique d'exploration de données (8 mn)
  - En synthèse, présentation des faits pertinents pour l'application (2 mn)
- **Discussion (5 minutes)**
- **Débriefing (5 minutes)**

Appel à projets et idée d'application

- **Appel à projets**
  - Objet: trouver des idées innovantes d'applications en lien avec l'alimentation
  - Basé sur les données d'Open Food Facts (version « française » de 2017)

- **Idée d'application**
  1. Sur la base des produits disponibles en France.
  2. L'utilisateur spécifie une demande combinant :
     - les **apports nutritionnels** contenant certaines **vitamines** et/ou sels **minéraux**
     - les **apports nutritionnels à minimiser** concernant le **sel**, le **sucre**, les **graisses saturées** ou les **graisses trans**
     - d'éventuels **éléments à exclure** en termes d'**allergènes**, d'**huile de palme**, et d'**additifs**.
  3. L'application propose en retour une **liste des meilleurs produits** classés selon leur qualité en termes de **score nutritionnel**.
  4. À partir d'un produit de la liste, l'utilisateur a la **possibilité d'élargir sa recherche** à d'autres produits en fonction de la **catégorie**.

Démarche méthodologique de nettoyage

1. **Chargement et prise de connaissance** du jeu de données
2. **Compréhension des caractéristiques** et structuration pour faciliter leur appréhension
3. **Premières constatations** sur le jeu de données et sur ses possibilités d'exploitation
4. Formulation d'une **idée d'application**
5. **Filtrage du jeu de données** sur les caractéristiques et données utiles à l'application
6. Suppression des **doublons**
7. Pour chaque caractéristique: définition, recherche et traitement des **valeurs aberrantes**
8. **Contrôles de cohérence** au niveau produit
9. Traitement des **manquants**

Chargement et prise de connaissance du jeu de données

- **162 variables** (colonnes)
  - Types: ‘object’ et ‘float64’
  - 16 variables sans aucune valeur → suppression : 146 variables
- **320 772 produits** (lignes)
- **76% de manquants** au total

Compréhension des caractéristiques

- Structuration arborescente des variables pour faciliter leur appréhension:
  - 2 niveaux de structure
  - Etat des manquants par variable
  - Type de variable (quantitative, catégorielle)

Structure des variables

- Variables de date / temps:
  - Variables '_t': conversion en numérique avec vérification d'erreur (coerce)
  - Autres variables: conversion au format 'datetime' avec vérification d'erreur (coerce)

- Variables de quantité: multiples unités → conservation telles qu'elles

- Variable de géolocalisation conservées telle qu'elle pour le moment

Premières constatations

☑ Certaines familles de variables sont suffisamment renseignées pour envisager leur exploitation par une application :

- Moins de 10% de manquants : ‘dates’, ‘keys’, ‘status’, ‘sale_location’, ‘product_name’, et ‘brands’.
- De 10% à 30% de manquants : ‘product_intake’ (‘energy’, ‘protein’, ‘salt’, ‘carbohydrates’, ‘fats’) et ‘composition’ (‘ingredients_text’, et les nombres concernant : ‘additives’ et ‘palm_oil’).
- De 30% à 40% de manquants : ‘nutrition_quality’.

☑ La famille ‘product_intake’ contient de nombreuses variable très peu renseignées, mais résulte probablement de leur faible présence dans la majorité des produits. Il en va de même pour les variables ‘composition’ concernant ‘allergens’, ‘traces’ et ‘palm_oil’.

☒ Avec plus de 88% de manquant, la famille de variable ‘traceability’ est difficilement exploitable.

⍰ La famille ‘product_category’ comporte de 70% à 75% de manquant pour 7 de ses variables, et 83.5% pour la 8ème . Cette information est particulièrement utile et nous examinerons plus avant sa meilleure utilisation possible.

A noter que la réduction du jeu de données à un pays impacte probablement ce taux

Filtrage du jeu de données en fonction de l'application

Réduction aux produit commercialisés au moins en France: (98 440) **et avec au moins une vitamine ou minéral:** 4042 produits et 46 variables

Tri des variables par nombre croissant de manquants pour prioriser les mieux renseignées

Examen des variables catégorielles:

- Liste des catégories et taille associées
- Nombre et taux de remplissage

Choix entre plusieurs variables catégorielles donnant des informations équivalentes:

→ Par comparaison des individus renseignées

- Si une variable renseigne au moins tous les individus renseigné par l'autre, elle peut être retenue en priorité
- Sinon, une variable renseigne les manquants de l'autre.

→ Par concaténation des valeurs des 2 variables pour chaque individu (valeurs séparées par un ';')

Choix des variables de catégorie de produit:

- Besoin d'une variable avec peu de catégories ('pnns_group_1')
- Complément naturel par 'pnns_group_2'

- Famille '**products**' :
  - Variables '**brands**' : conservation du meilleur des 2 variables en conservant la valeur de 'brands' lorsqu'elle existe et supprimant les valeurs identifiées comme NaN
  - Variables '**keys**' : nous ne conservons que la variable 'url'
  - Variables '**product**' : conservation de la variable 'product_name' ('image_url' est couvert par 'url')
  - Variables '**sale_location**' : nous conservons la variable 'stores', plus utile que 'purchase_places'
- Famille '**composition**':
  - Variables 'additives' : conservation des variables 'additives_n' et 'additives_tags', cette dernière plus explicite que 'additives_fr'.
  - Variables 'allergens' : conservation de 'allergens' (l'autre n'est pas renseignée)
  - Variables '...palm_oil...' : nous ne conservons qu'une variable de synthèse 'palm' indiquant si le produit contient ou pourrait contenir de l'huile de palme sous forme d'un booléen en conservant les NaN (équivalent à 3 valeurs: 'oui', 'non' et 'ne sait pas')
  - Variables 'traces' : nous ne conservons pas ces variables.
- Famille '**product_category**' : nous conservons les variables 'pnns_groups_1' et 'pnns_groups_2'.
- Famille '**product_intake**' :
  - Nous conservons les variables des familles 'vitamins' et 'minerals'
  - Sélection de variables clés concernant l'apport nutritionnel à minimiser: 'salt_100g', 'saturated-fat_100g', 'trans-fat_100g', et 'sugars_100g'.
- Famille '**quality**' : nous conservons les variables 'nutrition_grade_fr' et 'nutrition-score-fr_100g
- **Familles de variables non conservées**: 'dates', 'packaging', 'quantity', 'status' et 'traceability'.

Suppression des doublons

Pas de doublons

Valeurs aberrantes

- Examen des statistiques des variables numériques: suppression d'une variable avec des valeurs toutes nulles

- Détermination des valeurs de seuils minimum et maximum pour chaque variable:
  - Additifs: conservations après vérification de la crédibilité des valeurs extrêmes
  - Vitamines et minéraux: appui sur l'annexe XIII du règlement (UE) No 1169/2011 du 25 octobre 2011 pour fixer les seuils min (fonction de la valeur significative) et seuil max (fonction de l'apport journalier)
  - Nutri-score: fixé par sa définition entre -15 et +40
  - Composants à minimiser:
    - Seuil min à 0
    - Seuil max fixé de manière itérative en fonction des produits au-dessus du seuil.
      Ex: sel, seuil max 50g, en dehors du sel, pour les cubes de bouillon.
  - Conclusion: 120 / 4042 valeurs aberrantes à traiter


Seuils

- Traitement des valeurs aberrantes (vitamines et sels minéraux):
  - Au-dessous du seuil: 0
  - Au-dessus du seuil: valeur moyenne de la variable sur les produits ayant une valeur non nulle

Contrôles de cohérence

☒ Nutri-score: correspondance avec ‘nutrition_grade_fr’
→ correction individuelle des erreurs (boissons)

☑ *Somme des variables de type ‘_100g’ < 100 g*
(indépendantes sauf bicarbonate/sel)

☑ Vérification du nombre d’additifs avec le nombre d’additifs listés

Traitement des manquants

- Bilan des manquants: 65%
- Affectation directe de valeur:
  - ‘Additive_n’, famille vitamines et minéraux, famille toMinimize: 0
  - Variables ‘allergens’ et ‘additives_tags’: chaine de caractère vide
  - Variables ‘palm’ et ‘stores’: ‘***unknown***’
- Estimation de valeurs manquantes
  - Recherche de valeurs dans le cas de faible nombre de manquants:
    - ‘product_name’: 11/12 corrigés, reste un produit inexistant conduisant à la suppression du produit
    - ‘brands’ : 8/9 corrigés, reste une valeur indéterminée fixée à ‘***unknown***’
  - Déduction par corrélation avec une autre variable: 8 manquants de ‘pnns_groups_1’ traités par corrélation avec ‘pnns_groups_2’.
- Estimation par technique de machine learning s’agissant de ‘nutrition_grade_fr’, puis estimation par la moyenne pour ‘nutrition-score-fr_100g’:
  - Estimation de ‘nutrition_grade_fr’ par k-NN ;
  - Estimation de ‘nutrition-score-fr_100g’ comme la moyenne des valeurs connues pour la valeur correspondante de ‘nutrition_grade_fr’

Etat du jeu de données après nettoyage

- 45 variables (caractéristiques)
  - **37 numériques**: nombre d'additifs, vitamines, minéraux, sel, graisses, sucre et nutri-score
  - **6 catégorielles**: marque, magasin, huile palme, catégorie, sous-catégorie, nutri-grade
  - **2 informatives** (objet): nom du produit et url
- 4041 produits, 14 vitamines et 15 minéraux
- Aucun manquant

Démarche méthodologique d'exploration de données

1. Analyse univariée

- Caractéristiques numériques
- Caractéristiques catégorielles

2. Analyse bivariée

- Caractéristiques numériques entre elles
- Caractéristiques mixtes numériques / catégorielles

3. Analyse multivariée

- ACP : corrélation des variables numériques entre elles
- ANOVA et tests statistiques associés (p-value)

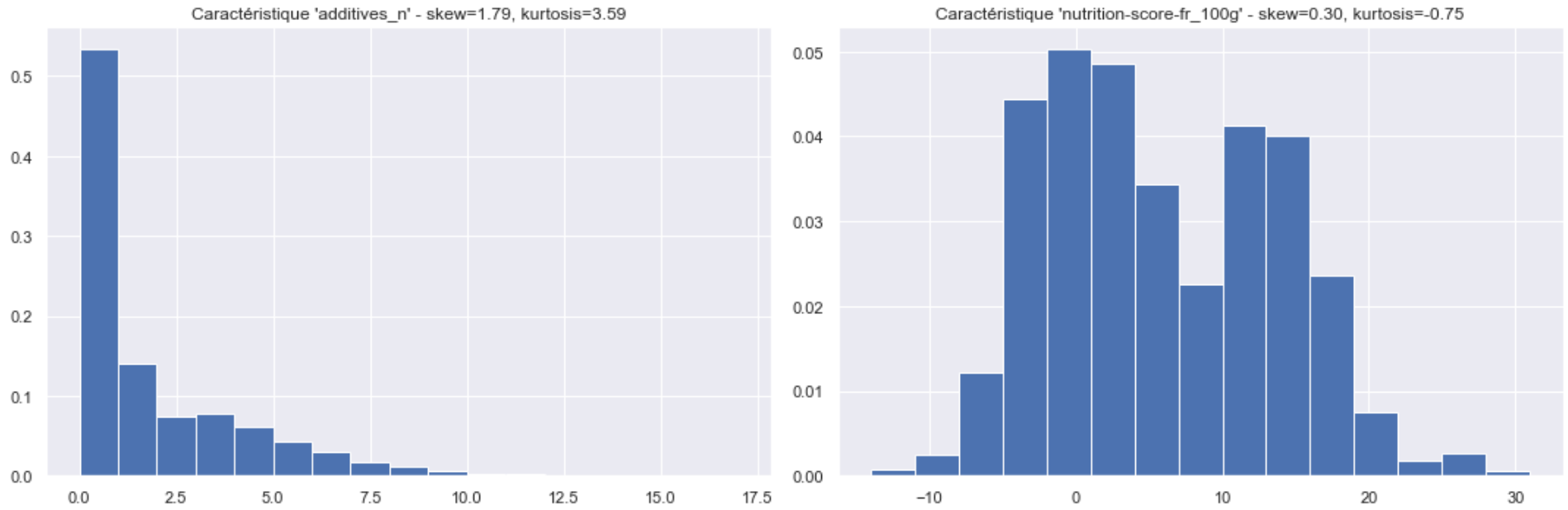
4. Faisabilité et pertinence de l'application

Analyse univariée - Caractéristiques numériques

- **Examen des caractéristiques statistiques**

	additives_n	nutrition-score-fr_100g
count	4041.000	4041.000
mean	1.496	5.395
std	2.219	7.769
min	0.000	-14.000
25%	0.000	0.000
50%	0.000	4.000
75%	3.000	12.000
max	17.000	31.000

- **Examen des distributions**



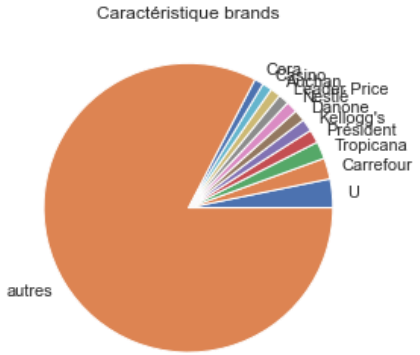
OPENCLASSROOMS
CentraleSupélec

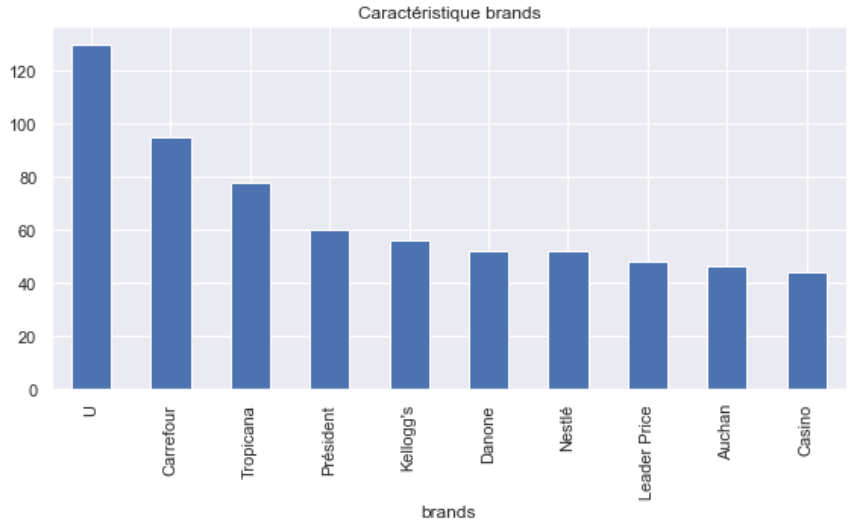
Analyse univariée - Caractéristiques catégorielles

- **Examen des caractéristiques statistiques**

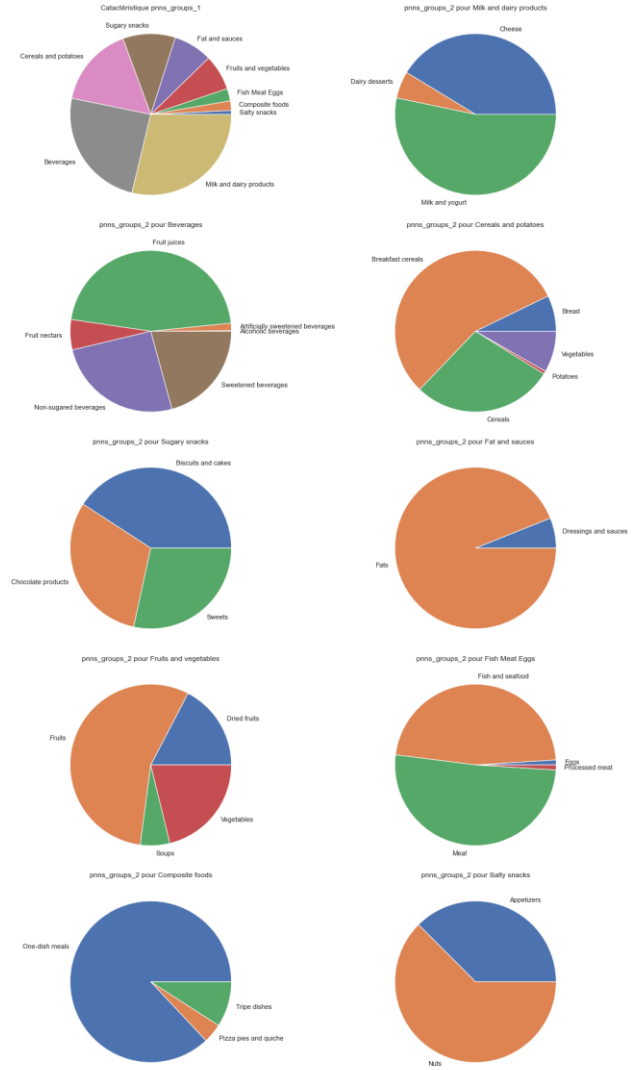
	palm	pnns_groups_1	pnns_groups_2	nutrition_grade_fr	brands	stores
count	4041	4041	4041	4041	4041	4041
unique	3	9	31	5	1381	414
top	False	Milk and dairy products	Milk and yogurt	a	U	***unknown***
freq	3179	1162	619	990	130	1398

- **Examen des fréquences et nombres de catégories**





Catégories et sous-catégories



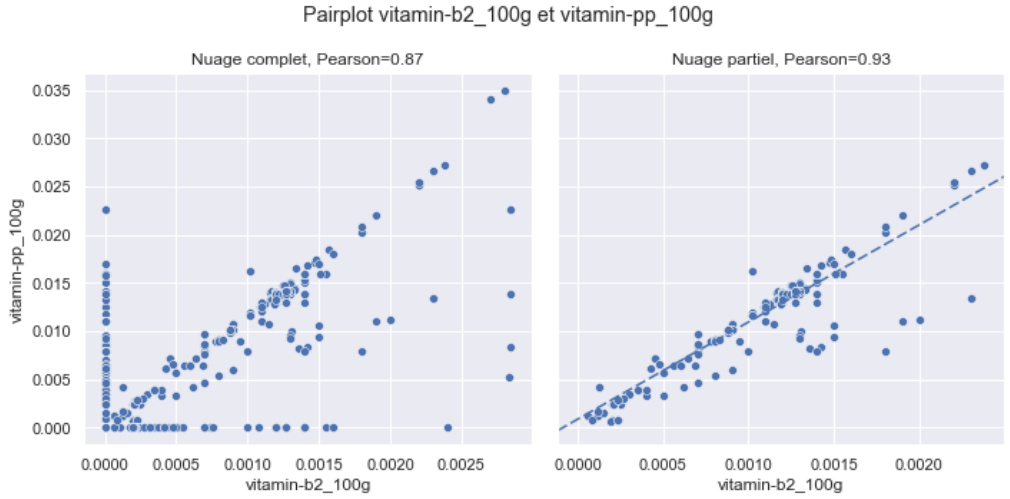
OPENCLASSROOMS
CentraleSupélec

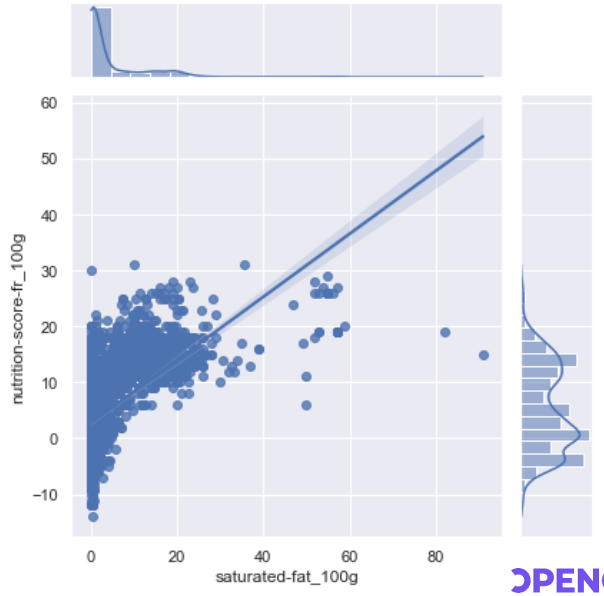
Analyse bivariée - Numériques entre elles

- **Recherche de corrélation, coef de Pearson > 0,55**

	vitamin-pp_100g	pantothenic-acid_100g	vitamin-b2_100g	silica_100g	bicarbonate_100g	nutrition-score-fr_100g	saturated-fat_100g
vitamin-pp_100g	0.00	0.63	0.87	-0.02	-0.02	0.14	-0.11
pantothenic-acid_100g	0.63	0.00	0.60	-0.01	-0.01	0.10	-0.08
vitamin-b2_100g	0.87	0.60	0.00	-0.02	-0.02	0.13	-0.09
silica_100g	-0.02	-0.01	-0.02	0.00	0.59	-0.06	-0.03
bicarbonate_100g	-0.02	-0.01	-0.02	0.59	0.00	-0.07	-0.04
nutrition-score-fr_100g	0.14	0.10	0.13	-0.06	-0.07	0.00	0.60
saturated-fat_100g	-0.11	-0.08	-0.09	-0.03	-0.04	0.60	0.00

- **Visualisation du nuage de points et régression linéaire**





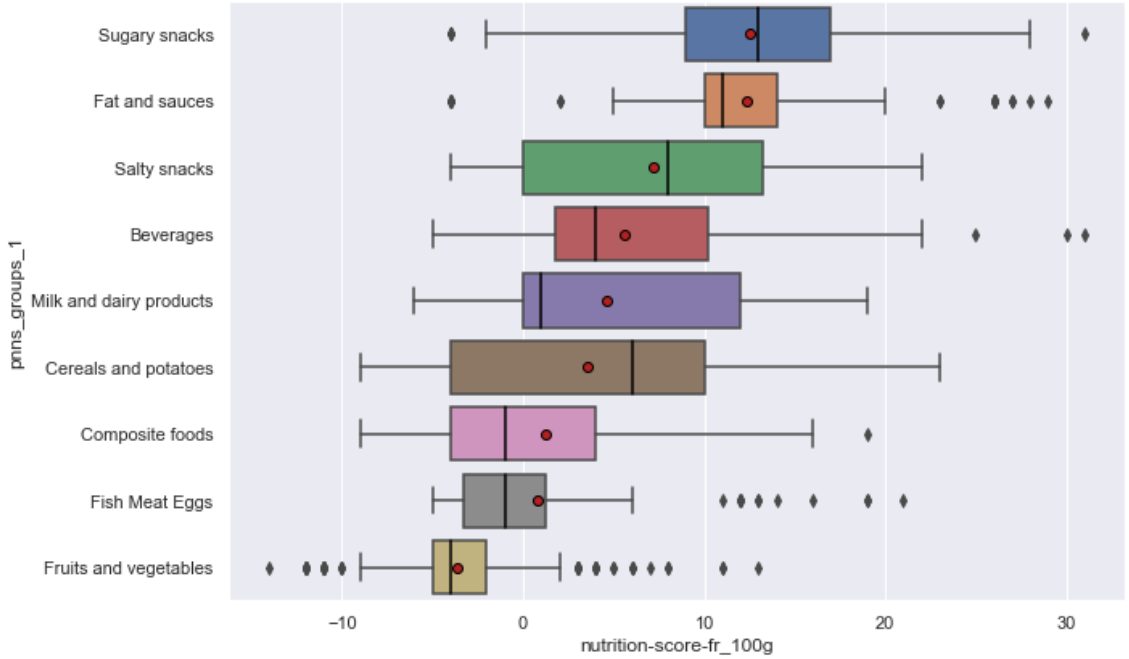
OPENCLASSROOMS
CentraleSupélec

Analyse bivariée - Numériques / catégorielles

- Recherche de corrélation, $\eta^2 > 0,5$

	calcium_100g	saturated-fat_100g	vitamin-pp_100g	nutrition-score-fr_100g	sugars_100g	vitamin-b2_100g
nutrition_grade_fr	0.13	0.29	0.04	0.88	0.09	0.04
pnns_groups_2	0.53	0.53	0.56	0.53	0.55	0.58

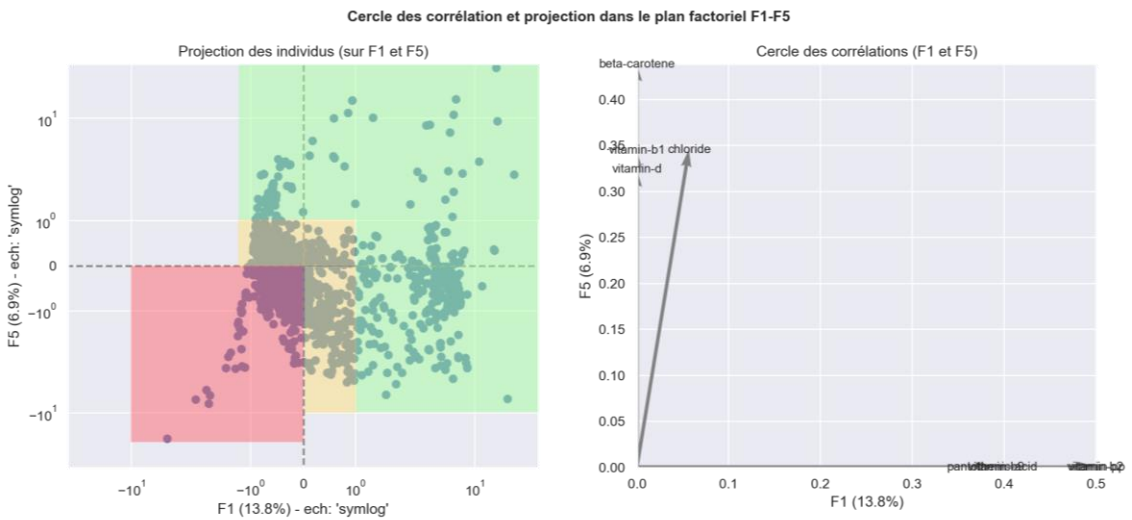
- Tracé de box-plot



Analyse multivariée - ACP

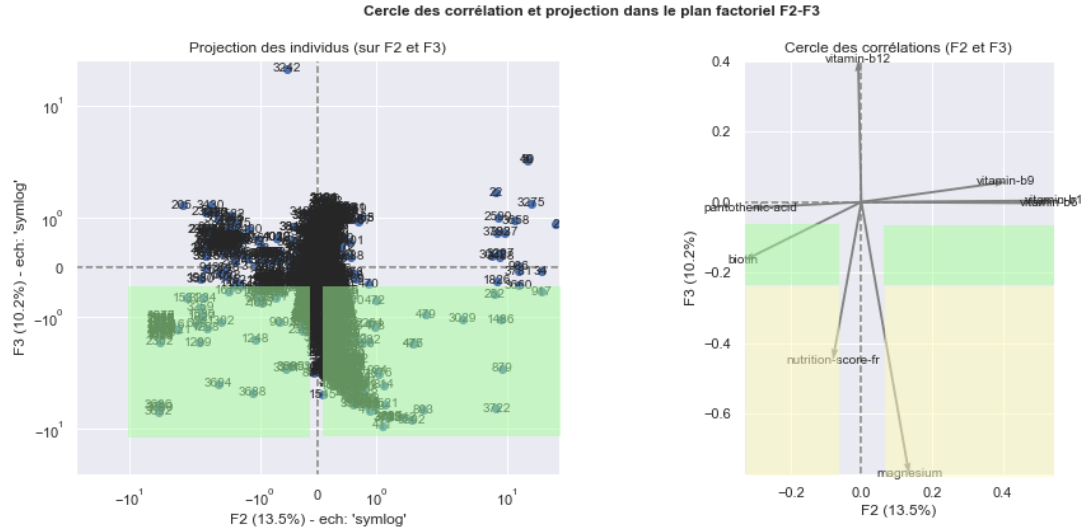
Analyse itérative, de l'ensemble des variables pour se raffiner progressivement :

- 35 variables, 4041 individus, 9 composantes (pour aller jusqu'au critère de Kaiser) ;
- réduction aux 22 variables (vecteurs projetés > 0.3) et 8 composantes ;
- sélection de 2 couples d'axes d'inertie : un plan pour les vitamines et un autre pour les minéraux.



Projection des individus:

- Analyse de couples (vitamine, minéral) à privilégier
- Recherche des meilleurs Nutri-score



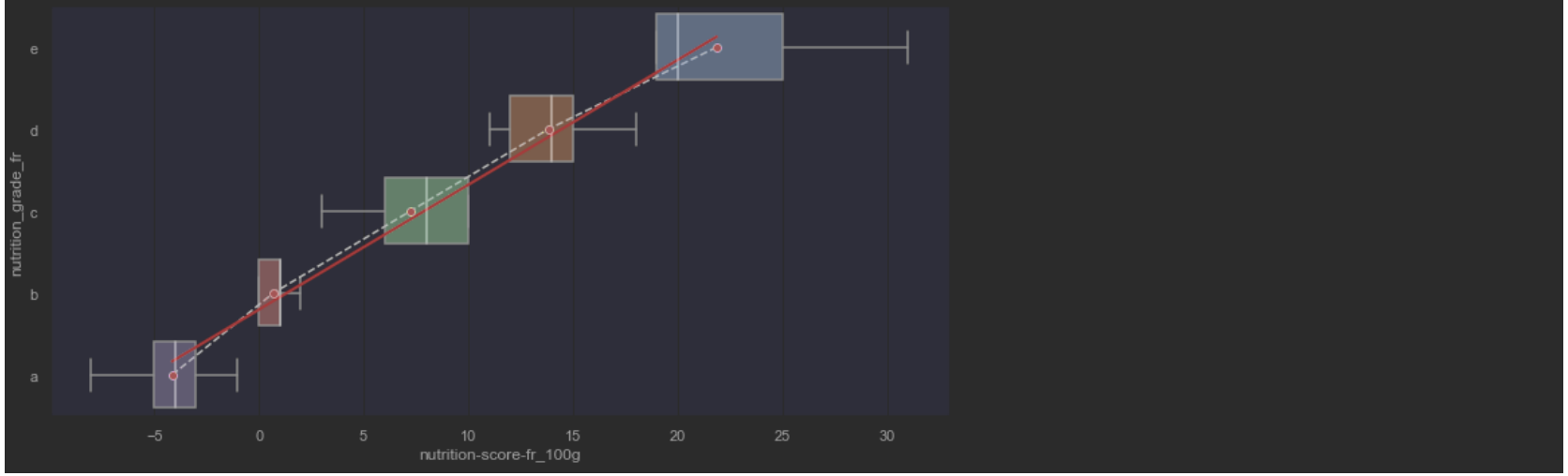
OPENCLASSROOMS
CentraleSupélec

Analyse multivariée – ANOVA et tests statistiques associés (1/2)

Exemple: nutri-score en fonction du nutri-grade pour la nourriture solide:

```
* ANOVA pour la nourriture solide
Rapport de corrélation pour les k= 5 catégories du graphique et n= 3053 données : η²=0.93
 -> Test de normalité de Shapiro positif pour toutes les catégories
 -> Test d'homoscédasticité de Bartlett négatif : p-value=3.20e-171
    Ecarts types: [3.071 2.141 2.457 0.769 2.219]
 -> Test de Welch (non égalité des moyennes) positif pour toutes les catégories
Résultat positif du test de Fisher : F=395.11 > 2.37  , et p-value=3.03e-85 < 0.05
 Rappel des hypothèses relatives au test :
  - H0 : les moyennes par catégories sont égales entre elles (les variables sont indépendantes)
  - H1 : la moyenne d'au moins une catégorie diffère des autres (les variables sont corrélées)

Moyenne catégorielle : 'nutrition-score-fr_100g' = -5.21  + 6.56 * 'nutrition_grade_fr', avec : 'a'= 0 , …, 'e'= 4
 -->Coefficient de corrélation r² =0.99
```



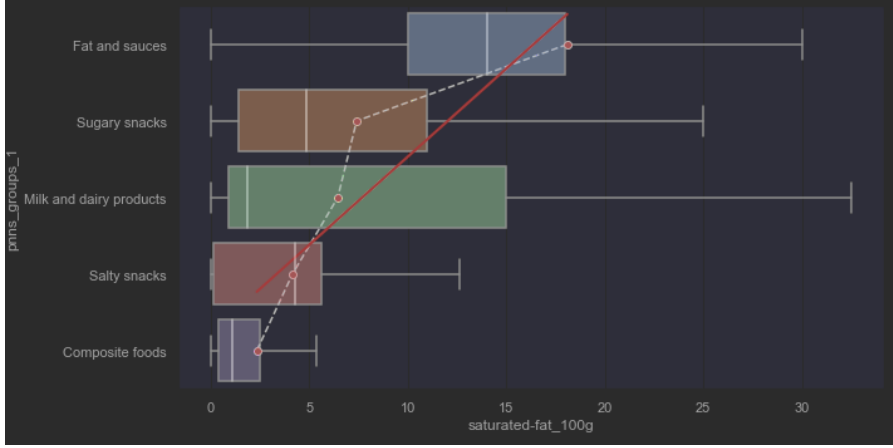
OPENCLASSROOMS
CentraleSupélec

Analyse multivariée – ANOVA et tests statistiques associés (2/2)

Exemples: graisses et sucres selon la catégorie de produit

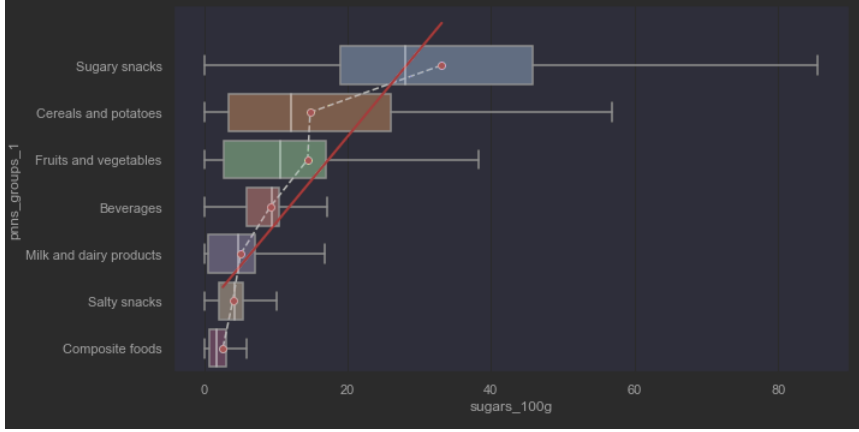
```
* ANOVA pour le taux de graisses saturées selon la catégorie du produit
Rapport de corrélation pour les k= 5 catégories du graphique et n= 2011 données : η²=0.18
 -> Test de normalité de Shapiro positif pour toutes les catégories
 -> Test d'homoscédasticité de Bartlett négatif : p-value=6.50e-86
    Ecarts types: [14.819  7.736  7.992  3.524  2.905]
 -> Test de Welch (non égalité des moyennes) positif pour toutes les catégories
Résultat positif du test de Fisher : F=871.73 > 2.38  , et p-value=1.41e-173 < 0.05
 Rappel des hypothèses relatives au test :
  - H0 : les moyennes par catégories sont égales entre elles (les variables sont indépendantes)
  - H1 : la moyenne d'au moins une catégorie diffère des autres (les variables sont corrélées)

Moyenne catégorielle : 'saturated-fat_100g' = -0.99  + 4.34 * 'pnns_groups_1', avec : 'Composite
foods'= 0 , …, 'Fat and sauces'= 4
 -->Coefficient de corrélation r² =0.80
```




```
* ANOVA pour le taux de sucre selon la catégorie du produit
Rapport de corrélation pour les k= 7 catégories du graphique et n= 3627 données : η²=0.35
 -> Test de normalité de Shapiro positif pour toutes les catégories
 -> Test d'homoscédasticité de Bartlett négatif : p-value=0.00e+00
    Ecarts types: [21.955 12.245 16.555  9.438  5.299  2.778  3.952]
 -> Test de Welch négatif entre les catégories ' Cereals and potatoes ' et ' Fruits and vegetables '
 : W=0.31, p-value=0.758, dof=430.96
Résultat positif du test de Fisher : F=1467.88 > 2.10  , et p-value=1.22e-292 < 0.05
 Rappel des hypothèses relatives au test :
  - H0 : les moyennes par catégories sont égales entre elles (les variables sont indépendantes)
  - H1 : la moyenne d'au moins une catégorie diffère des autres (les variables sont corrélées)

Moyenne catégorielle : 'sugars_100g' = -4.44  + 5.45 * 'pnns_groups_1', avec : 'Composite foods'= 0 ,
 …, 'Sugary snacks'= 6
 -->Coefficient de corrélation r² =0.80
```



OPENCLASSROOMS CentraleSupélec

Conclusion sur la faisabilité et pertinence de l'application

- **Faisabilité:**
  - **Une population de produits significative contenant des vitamines et minéraux: 4041**
  - **Une variété significative de vitamines (14) et de minéraux (15)**

- **Pertinence:**
  - **Des couples intéressants de vitamines et minéraux soutenus par une population significative**
  - **En classant les résultats par Nutri-score, cela conduit automatiquement à présenter un taux faible de graisses saturées (corrélation des graisses saturées avec le nutri-score)**
  - **Dans son choix de catégorie de produits, l'utilisateur peut écarter les catégories à fort taux de graisses saturées ou de sucres.**

- **Maquette de l'application: exemple avec le couple (vitamine B1, magnésium)**

product_name	url	brands	stores	nutritio	pnns_groups_1	additives_n	palm	saturated-fat_100g	vitamin-b1_100g	magnesium_100g
Mogette de Vendée Sabarot	http://world-	Sabarot	Monoprix	a	Fruits and vegetables	0.0	False	0.00	5.000e-04	0.180
Haricots cocos Sabarot	http://world-	Sabarot	Carrefour	a	Fruits and vegetables	0.0	False	0.00	5.000e-04	0.180
Pois cassés de France Sabarot	http://world-	Sabarot	Carrefour	a	Fruits and vegetables	0.0	False	0.00	5.000e-04	0.089
Noix de Grenoble, sèches	http://world-	Valnoix,Ville SA	Super U	a	Salty snacks	0.0	False	0.00	2.800e-04	0.141
Pois Chiches	http://world-	La Vie Claire	***unknown***	a	Cereals and potatoes	0.0	False	0.00	5.000e-04	0.115
Lentilles Corail	http://world-	La Vie Claire	La Vie Claire	a	Cereals and potatoes	0.0	False	0.00	5.000e-04	0.107
Dattes de Tunisie	http://world-	Nouri	Dia	a	Fruits and vegetables	0.0	***unknown***	0.00	3.962e-01	0.063
Hydra Ananas Coco	http://world-	Aptonia	Décathlon	a	Beverages	7.0	False	0.10	8.400e-05	0.028
haricots rouges	http://world-	Vivien Paille	Auchan	a	Cereals and potatoes	0.0	False	0.15	6.500e-04	0.138
Lentilles Vertes	http://world-	Vivien Paille	***unknown***	a	Fruits and vegetables	0.0	False	0.20	5.000e-04	0.100

Echanges avec l’évaluatrice

- **Discussion**
- **Débriefing**

OPENCLASSROOMS CentraleSupélec

Contact: Eric TREGOAT

eric.tregoat@gmail.com

06 49 99 79 59

https://www.linkedin.com/in/erictregoat/