

SOUTENANCE PROJET 3



Etude et faisabilité d'une application au
service de la santé liée à l'alimentation



ORDRE DU JOUR

INTRODUCTION

- ⇒ Problématique et appel à projets
- ⇒ Idée d'application



NETTOYAGE DES DONNÉES

- ⇒ Présentation de la base de données
- ⇒ Sélection des données pertinentes
- ⇒ Traitement données aberrantes et manquantes

EXPLORATION DES DONNÉES

- ⇒ Analyse des Nutriscores par catégories avec χ_2
- ⇒ ACP sur les plats préparés et corrélations entre les variables
- ⇒ Mise en place de l'étude d'un régime spécifique avec ANOVA



CONCLUSION

- ⇒ Synthèse travail effectué
- ⇒ Conclusion sur l'idée
- ⇒ Aller plus loin

INTRODUCTION

⇒ Problématique et appel à projets

⇒ Idée d'application

Problématique et appel à projets



➤ Santé Publique France:

- Trouver des idées innovantes d'applications en lien avec l'alimentation
- Open Food Fact, jeu de données composé d'informations générales, de compositions et nutritionnelles de nombreux aliments sur le marché

➤ Problématique:

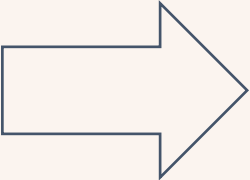
- Analyser la base de données
- Rechercher une idée d'application
- Etudier la faisabilité de l'idée

Idée d'application

Road2Health



- ❖ Il y a beaucoup de pathologies qui nécessitent un régime spécifique et contraignant
- ❖ A la sortie de l'hôpital, un patient se retrouve chez lui avec une fiche de conseils, contraintes...

- 
- ❖ Aide aux patients et assistance aux médecins
 - ❖ Conseils sur les aliments, orienter l'alimentation selon des restrictions, créer des programmes alimentaires...



NETTOYAGE DES DONNÉES:

- ⇒Présentation de la base de données
- ⇒Sélection des données pertinentes
- ⇒Remplissage des valeurs manquantes

Présentation de la base de données

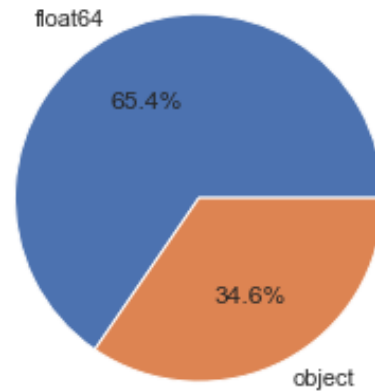


162 colonnes (attributs) valeurs nutritionnelles , info produit...

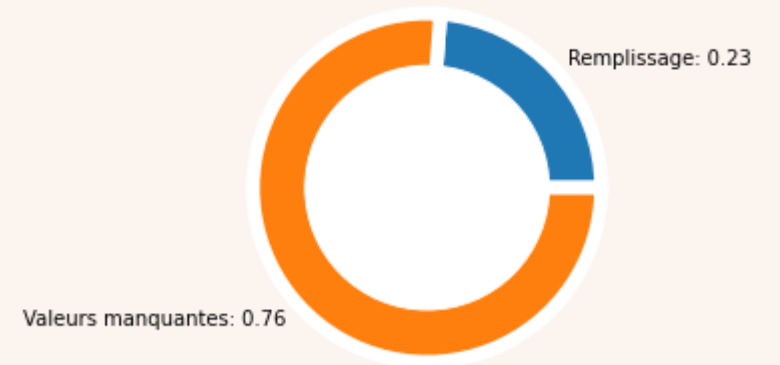
320 772
individus

Tout type d'aliment

Pieplot des types de variables dans la bdd



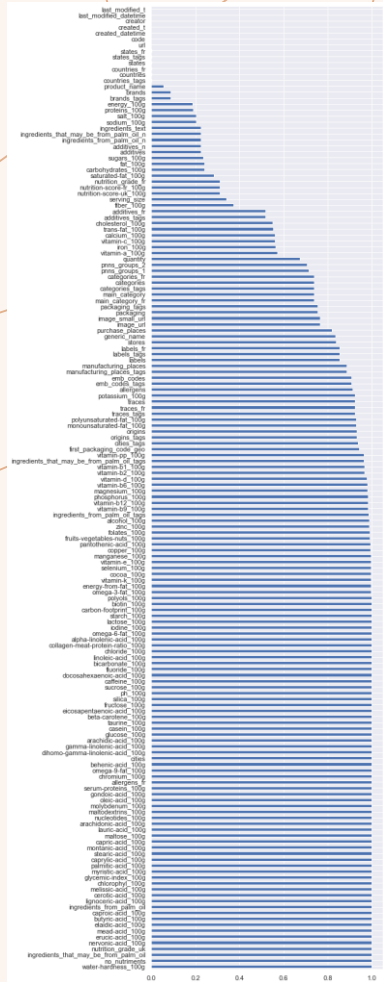
Remplissage de la base de données sans modifications



Nettoyage des données

Sélection des données pertinentes

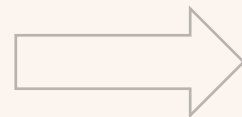
Taux de valeurs manquantes par variables



1) J'ai gardé seulement les variables qui avaient un taux de valeurs manquantes inférieurs à 75%: 162 -> 50 variables.
(vérifications pertinence de la perte.)

2) Variables pertinentes gardées pour l'étude:
Nutrition_grade_fr, main_category_fr, ingredients_texts, code, iron_100g, sodium_100g, fat_100g, protein_100g, fiber_100g, calcium_100g, carbohydrates_100g, energy_100g, quantity

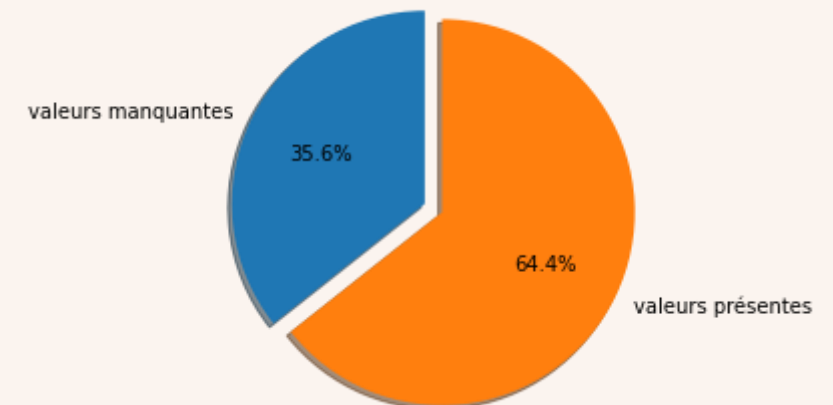
Nettoyage des données



320 772
variables

12 variables

Pie Plot Valeurs présentes/manquantes

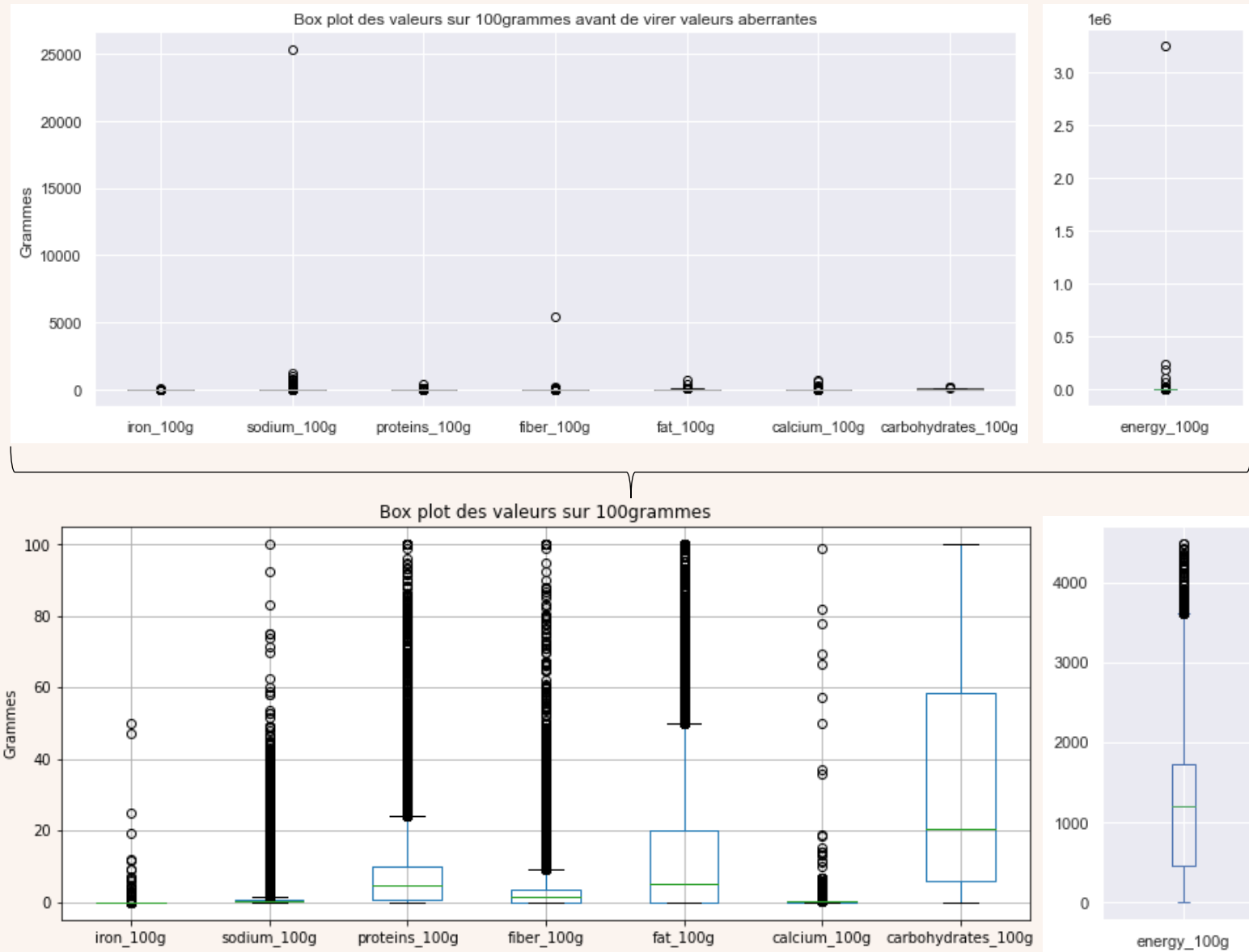


Remplissage des valeurs manquantes

Gestion données aberrantes:

- Variables Grammes:
 - $\in [0 ; 100]$
 - $48 \notin [0 ; 100] \rightarrow \text{NaN}$
- Variable energy_100g:
 - KJ (mais...)
 - Seuil à 4500
→ 277 individus supprimés

Nettoyage des données



Remplissage des valeurs manquantes

Choix pour nettoyage:

- Se limiter aux individus dont on connaît le nutriscore:

→ 320 771 -> 220 933 individus

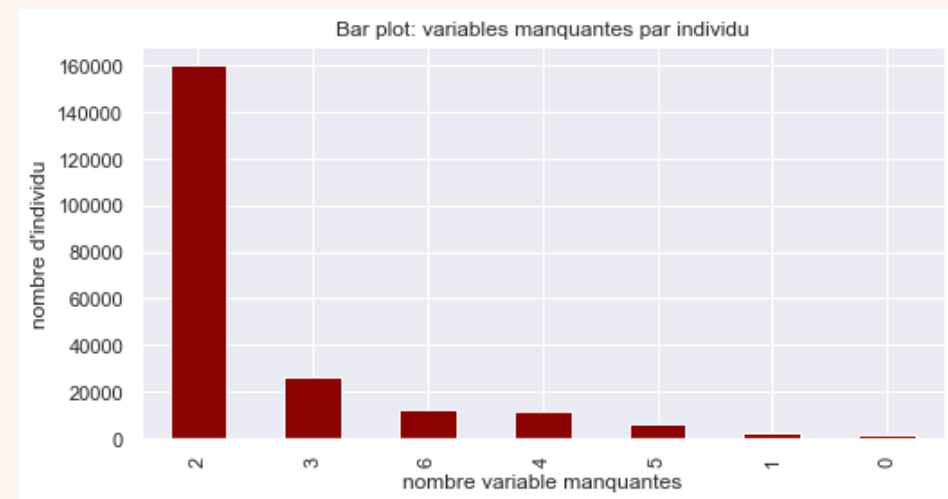
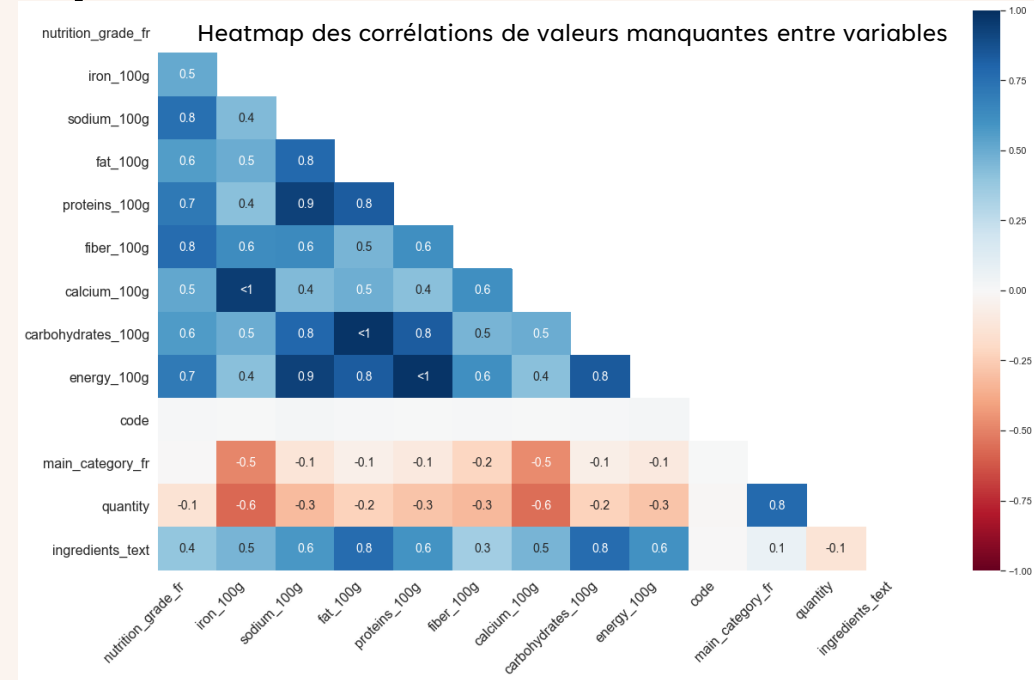
- Supprimer les individus avec +3 valeurs manquantes:

→ 220 933 -> 190 260 individus
(-30673 individus)

- Garder les valeurs manquantes de:

- ingredients_text
- main_category_fr
- quantity

Nettoyage des données



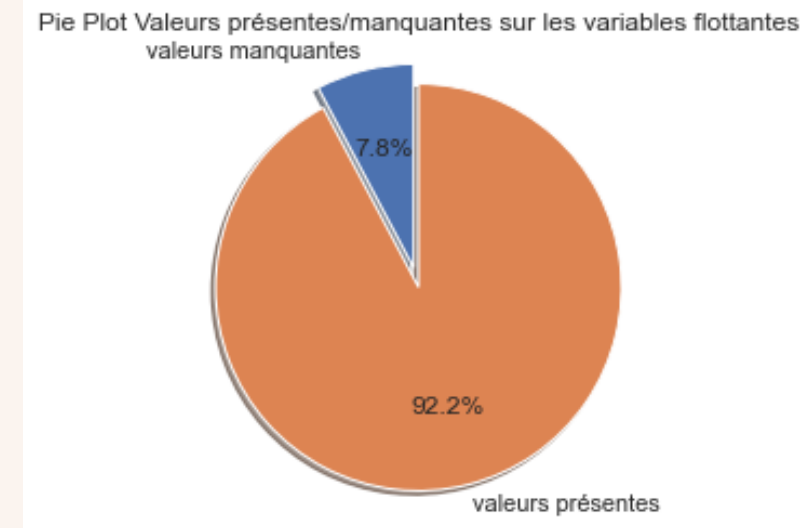
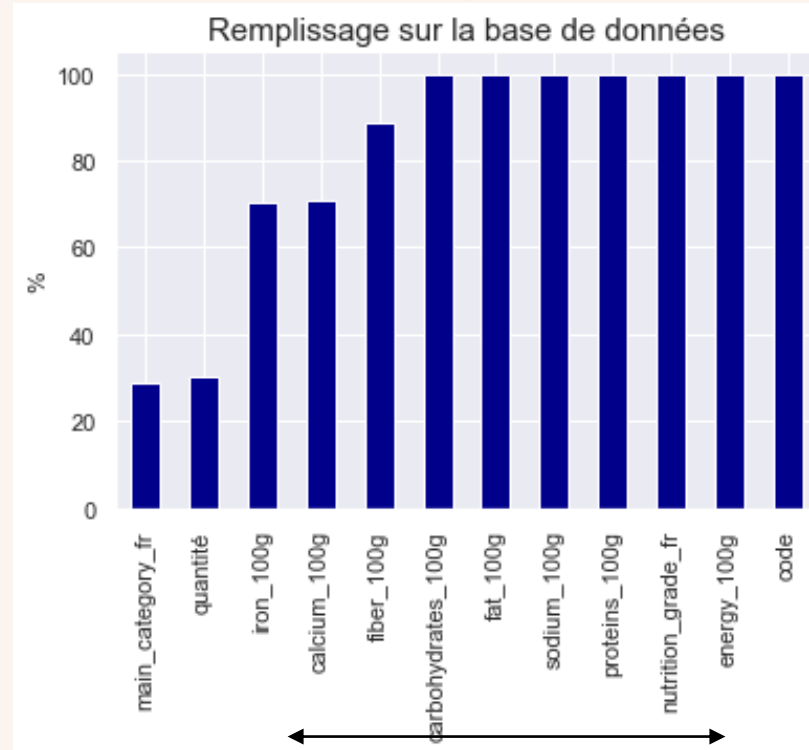
Remplissage des valeurs manquantes

Rappel état du Dataset:

- 12 columns, 190260 individus
- 133230 valeurs manquantes

⇒ Méthodes de remplissage utilisées:

- ✓ Moyenne
- ✓ KNN imputer de Scikit-Learn
- ✓ Moyenne par classe de NutriScore
- ✓ Moyenne par classe d'un K-means



Remplissage des valeurs manquantes

⇒ Moyenne

iron_100g	0.003080
sodium_100g	0.479138
fat_100g	13.397682
proteins_100g	7.871833
fiber_100g	2.881910
calcium_100g	0.097599
carbohydrates_100g	33.168061
energy_100g	1180.585248

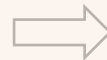


Remplissage par valeurs moyenne.

Pros: Rapide

Cons: Très simple et naïf

⇒ KNN imputer de Scikit-Learn



K=2

Pros: Automatique et proposé par une librairie

Cons: Très lent~16 minutes

Remplissage des valeurs manquantes

⇒ Moyenne par classe de NutriScore

	iron_100g	sodium_100g	proteins_100g	fat_100g	fiber_100g	calcium_100g	carbohydrates_100g	energy_100g
nutrition_grade_fr								
A	0.004431	0.131887	8.227365	2.620733	4.524347	0.064682	28.499851	707.317254
B	0.001107	0.215466	5.160853	3.961888	1.753193	0.054118	17.631122	523.771864
C	0.002978	0.666530	7.096964	10.131754	3.532765	0.077739	32.026005	1025.088574
D	0.004461	0.631003	8.774654	18.021554	2.765671	0.088912	40.622642	1491.026235
E	0.001595	0.568538	9.202548	26.984099	1.658690	0.193938	39.820774	1824.052733

Remplissage avec la moyenne pour chaque classe du Nutriscore
à l'aide de 2 fonctions

Pros: Meilleure précision pour remplissage des données.

Cons: On peut pas l'appliquer sans connaître le Nutriscore

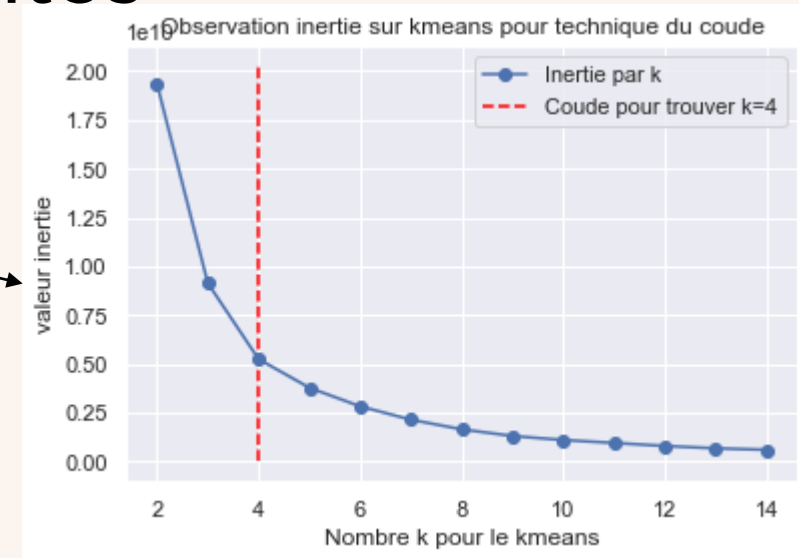
Remplissage des valeurs manquantes

⇒ Moyenne par classe d'un K-means

Entraînement de plusieurs Kmeans pour trouver le bon K:

Remplissage avec la moyenne de la classe correspondante

	iron_100g	sodium_100g	proteins_100g	fiber_100g	calcium_100g	fat_100g	carbohydrates_100g	energy_100g
classkM								
0	0.002354	0.619218	8.654562	2.267301	0.098870	9.847065	29.310595	987.076109
1	0.003499	0.393488	10.271907	4.996022	0.079145	35.751817	46.747146	2256.420792
2	0.000901	0.506329	4.023243	1.282019	0.044115	1.634669	10.456820	296.728818
3	0.004973	0.527755	9.988238	3.493923	0.147549	12.765816	57.067177	1587.592359

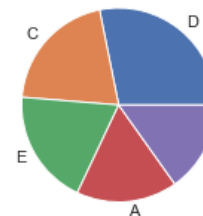


Pros: Précision intéressante car remplissement par clustering

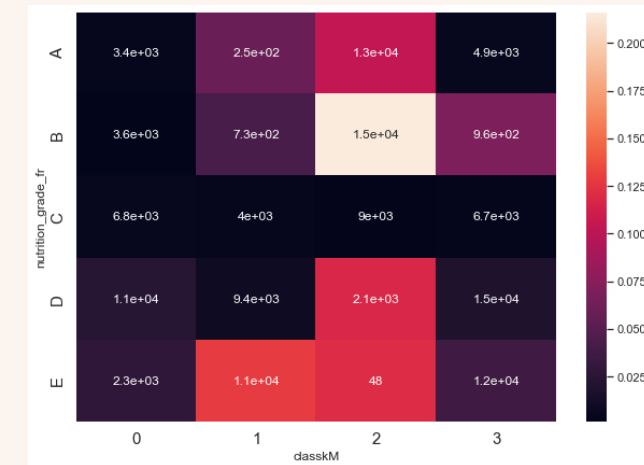
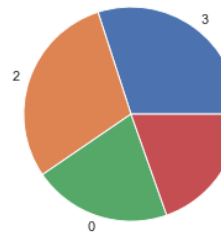
Cons: Pas précis s'il manque beaucoup de valeurs

Observations corrélation ClasseKm et Nutriscore

Pie Plot de la distribution des aliments par nutriscore (A-E)



Distribution des 4 clusters Du Kmeans



Nettoyage des données

Test hypothèse de dépendance du χ^2
→ Rejeté donc il existe une dépendance

EXPLORATION DES DONNÉES:

⇒ Test d'indépendance du χ^2 : Nutriscore / Catégorie

⇒ PCA sur les données: Ressem

⇒ Anova qualitative/quantitative: dépendance Nutri-Score/Fibre ?

Test d'indépendance du χ^2 : Nutri-Score / Catégorie

Nutri-Score:

- Indicateur de qualité nutritionnelle d'un aliment
- Il est lié aux valeurs nutritionnelles (sodium, glucide,...)

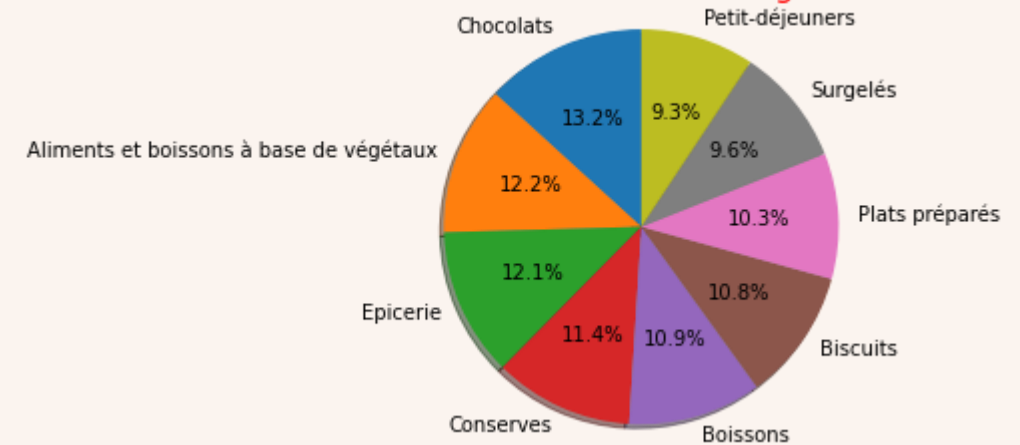
*Question dans l'objectif d'un régime spécial:
Est-ce qu'il y a une différence de distribution du Nutri-Score entre les différentes catégories ?*

➡ Test d'indépendance du χ^2 sur 16 376 individus

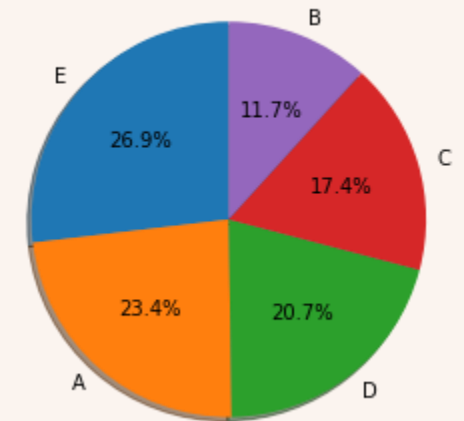
- H_0 : Elles sont indépendantes
- H_a : Elles ne sont pas indépendantes

Exploration des données

Pie Plot Différentes catégories



Pie Plot Différents Nutriscore



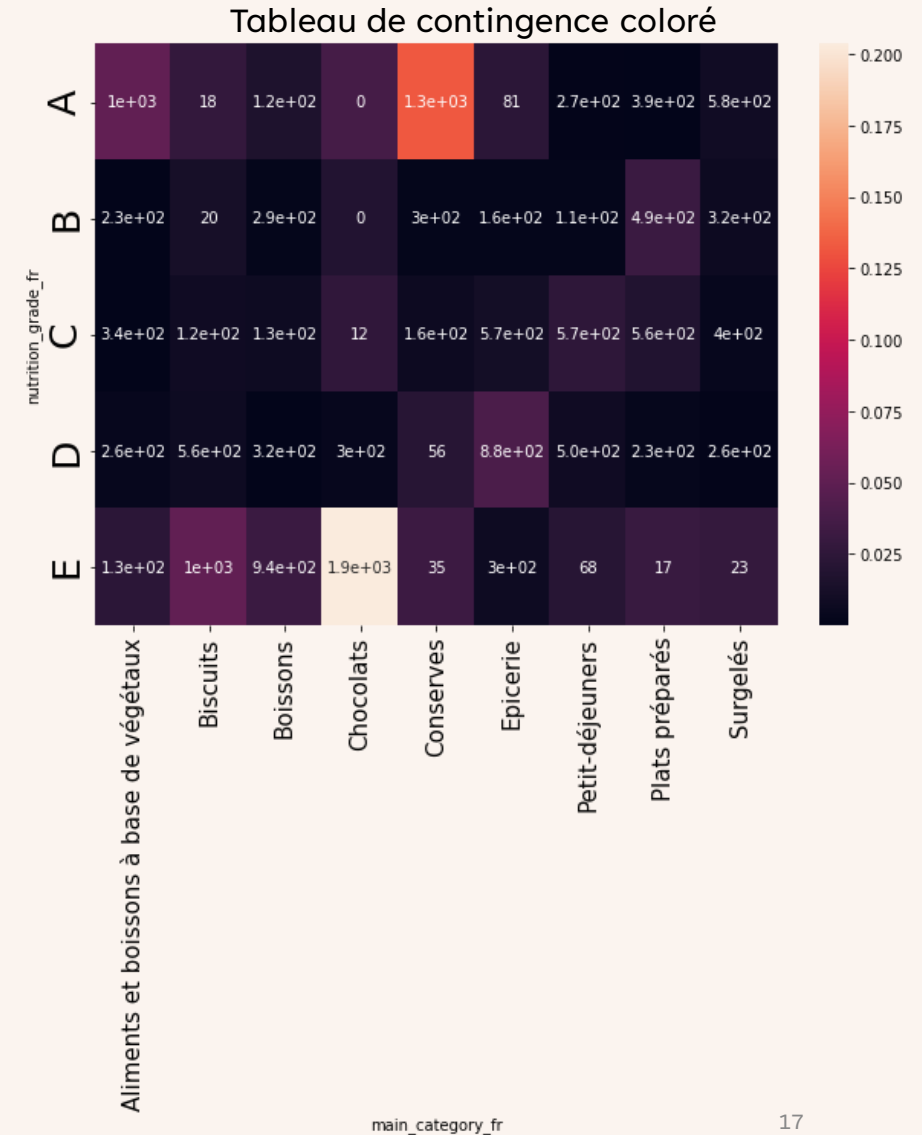
Test d'indépendance du χ^2 : Nutri-Score / Catégorie

Remarques:

1. On peut remarquer que les chocolats sont généralement des E
2. Les conserves plus vers des A

Test Statistique:

- $\chi^2 = 13589.42$
 - p-value = 0.0
- Hypothèse d'indépendance rejetée



Analyse en composantes principales

Objectifs:

- Liaison entre les variables ?
- Ressemblance/Différence entre individus ?

⇒ Peut-on faire une autre approche qu'avec les valeurs nutritionnelles classiques ?

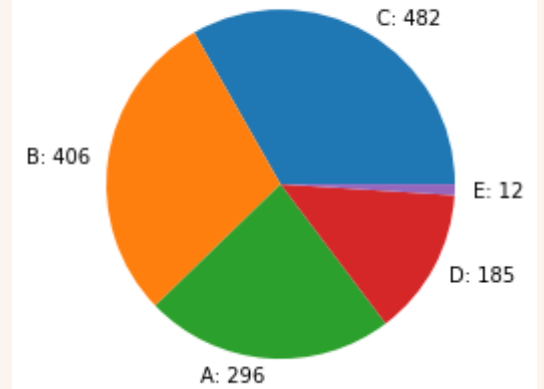
Pour l'étude: On se concentre sur les plats préparés:

- Conversion et/ou sélections des quantités en grammes
- 1381 individus

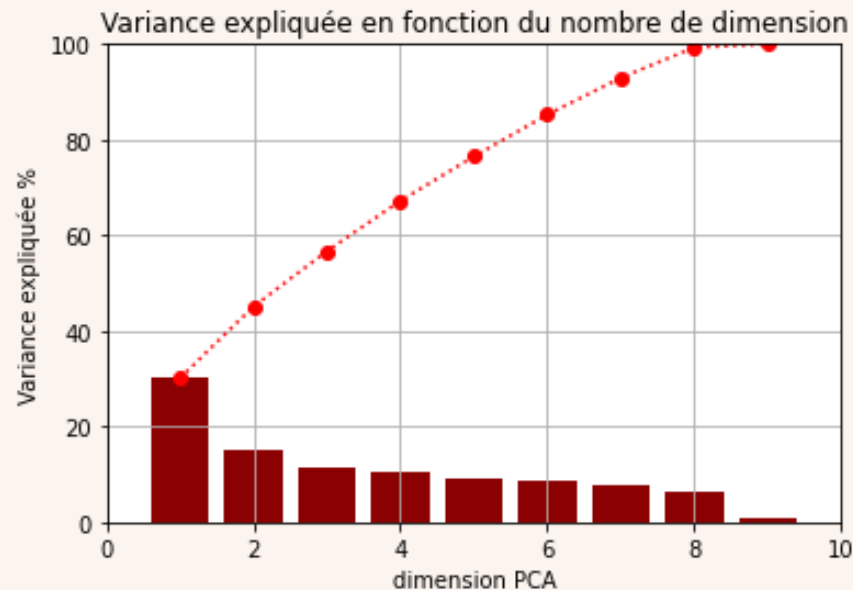
ACP sur données centrées/réduites:

Exploration des données

Distribution plat préparés par nutriscore



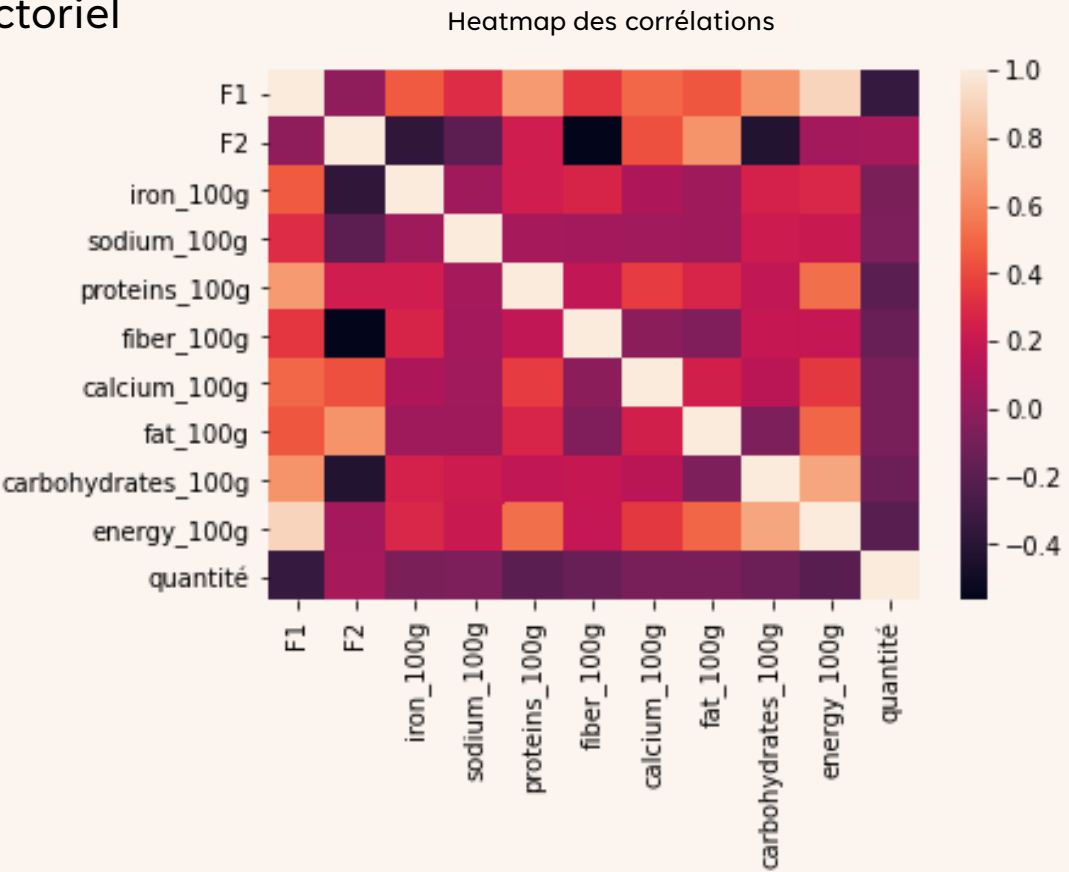
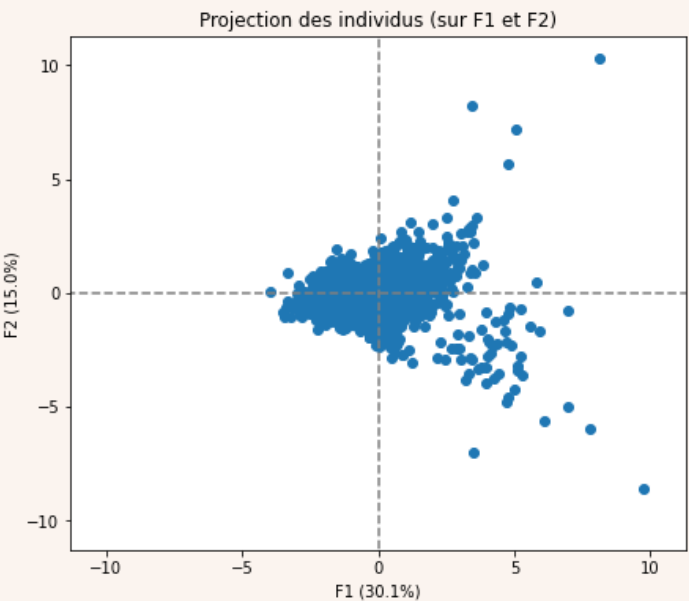
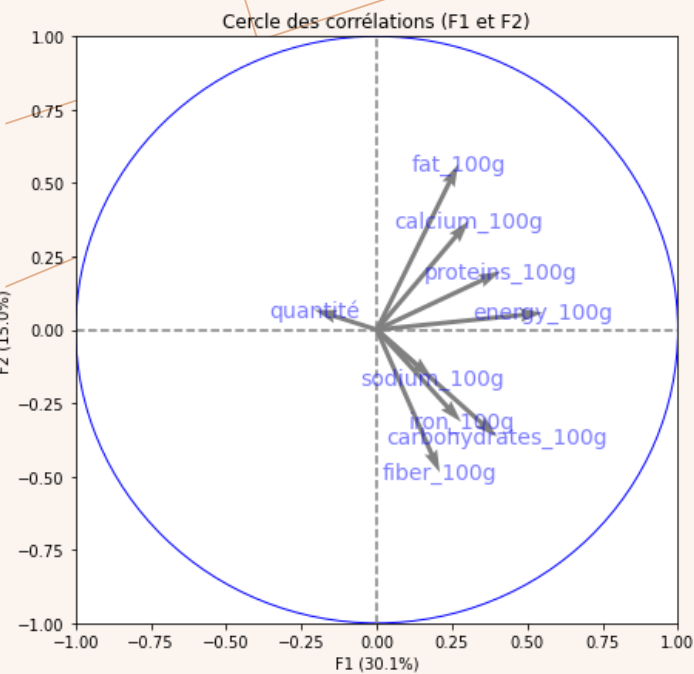
['iron_100g','sodium_100g','proteins_100g',
'fiber_100g','calcium_100g','fat_100g',
'carbohydrates_100g','energy_100g','quantité']



Analyse en composantes principales

Observations:

- Pas de variable clairement représenté sur plan 1^{er} factoriel
- Energie, protéine sont corrélées à F1
- Peut-être valeurs atypiques

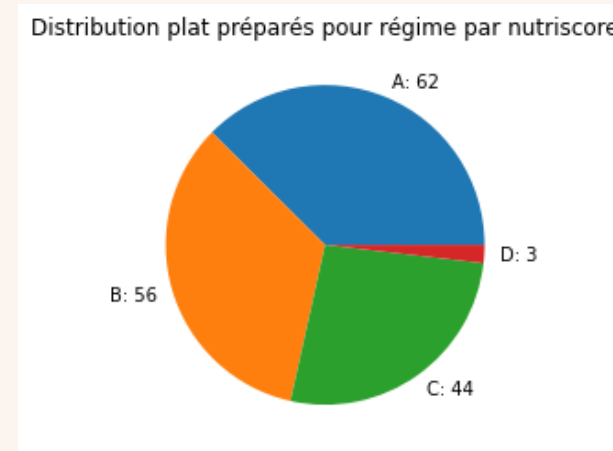


	energy_100g	proteins_100g	carbohydrates_100g	calcium_100g	iron_100g	fat_100g	fiber_100g	sodium_100g	quantité
F1	0.551659	0.411825	0.399506	0.306184	0.280065	0.270847	0.209416	0.184145	-0.207748

⇒ Anova Variables Fibres/Nutri-Score

- Les variables sont changées en pourcentage d'apport journalier (35% pour 1 repas...)
- Les plats préparés sont triés selon un régime pour un patient atteint d'une greffe de poumons (en gardant les données en fibre et Nutri-Score)
- Des informations sur la classification:

165 plats répondant aux contraintes sur 1381



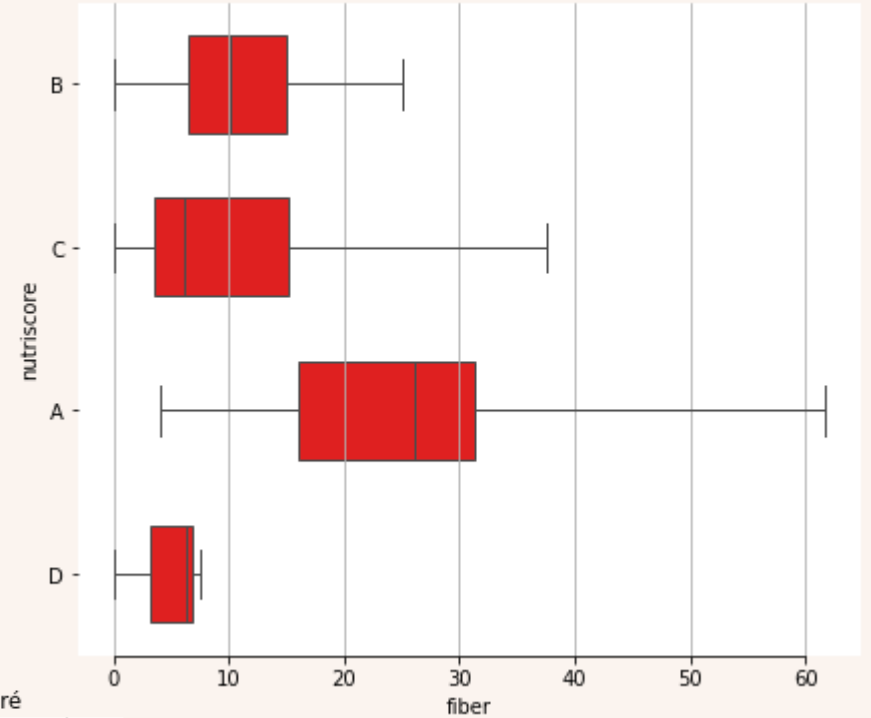
- Question: Selon l'apport journalier en fibre, a-t-on une distribution similaire pour les Nutri-Score ?

⇒ Anova Variables Fibres/Nutri-Score

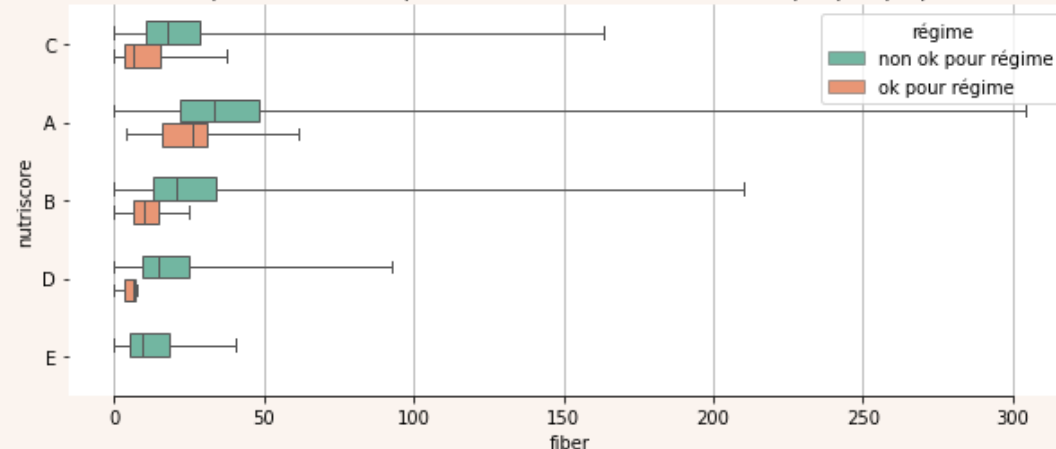
Observations:

- Il y a plus plats préparés noté A.
- On obtient un η^2 à 0,37 donc on pourrait supposer une différence de distribution de fibre pour les Nutri-Scores
- Finalement $p\text{-value} \approx 5.e^{-16}$ donc on peut rejeter l'hypothèse d'indépendance
- Remarque: On observe la même tendance de comportement sur les plats qui ne sont pas dans le régime.

Boxplot du % valeur quotidienne de fibre et le nutriscore par plat préparé

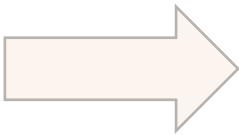


Boxplot du % valeur quotidienne de fibre et le nutriscore par plat préparé



CONCLUSION: SYNTHÈSE

- ✓ Base de données assez peu remplie → traitement, choix, sélection et un remplissage des données manquantes
- ✓ Etude: la catégorie et le Nutri-Score ne sont pas indépendant
- ✓ Les variables sont assez bien représentatives: on ne peut pas diminuer la dimension
- ✓ Cas particulier d'un régime: Il y a une dépendance entre le taux de fibres et Nutri-Score



Confirme une faisabilité d'exploitation de la base de données Open Food Fact pour une application de santé

CONCLUSION: ALLER PLUS LOIN

- ⇒ Evaluer et rechercher le remplissage le plus précis
- ⇒ Faire des analyses sur d'autres catégories d'aliments (surgelé, conserve, boisson...)
- ⇒ Scraper sur internet des recettes (Marmiton) et calculer l'apport pour une évaluations sur un régime

A series of thin, light brown lines crisscrossing the left side of the slide, creating a complex, abstract geometric pattern.

MERCI!

QUESTIONS ?

