Exploration et visualisation des données

- Démarche méthodologique de nettoyage
- Démarche de l'exploration de données

Analyse basée sur le jeu de données Open Food Facts

openfoodfacts.org

Environnement de travail

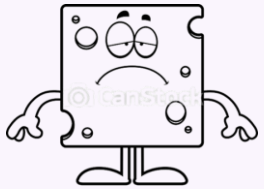
Librairies python spécialisées importées :

- Pandas
- Matplotlib
- Numpy
- Seaborn

Démarche méthodologique de nettoyage

1 932 482 lignes (produits) et **186 colonnes** (variables)

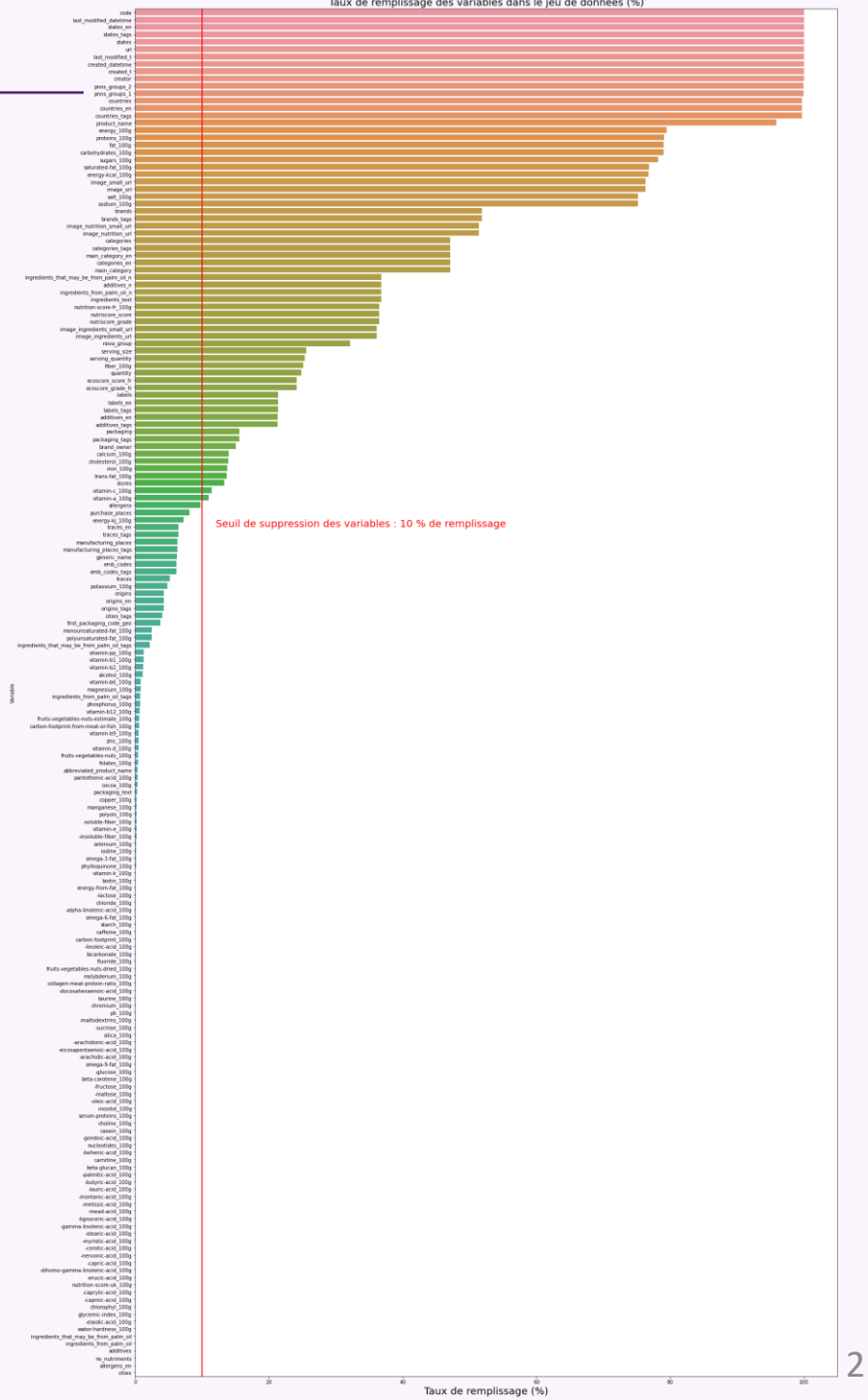
Cellules avec des données manquantes : 80 %
Lignes avec au moins une donnée manquante : 100 %



1 932 482 Produits - **186** Variables
(**126** quantitatives , **60** qualitatives)

Traitement des données

1 527 078 Produits - **37** Variables
(25 quantitatives , **12** qualitatives)

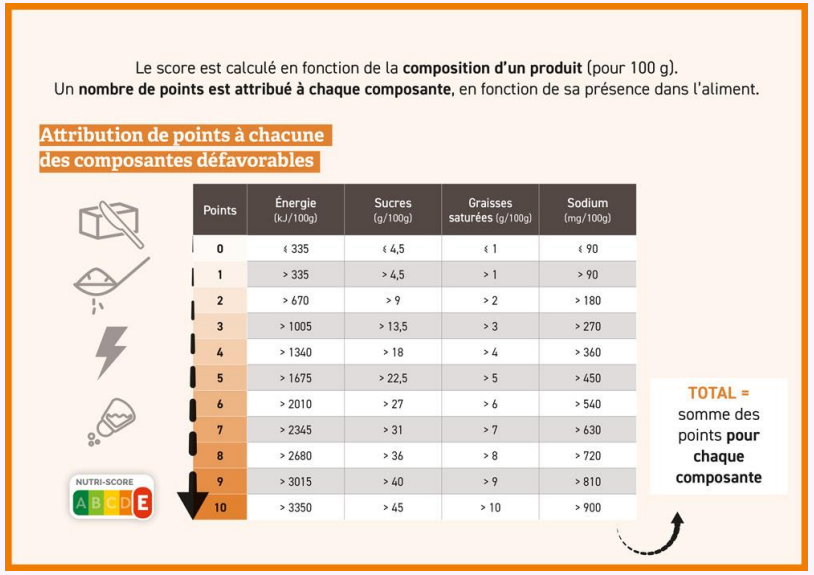


Démarche méthodologique de nettoyage

Détection des valeurs impossibles

Imputation des valeurs manquantes

Calcul du nutriscore

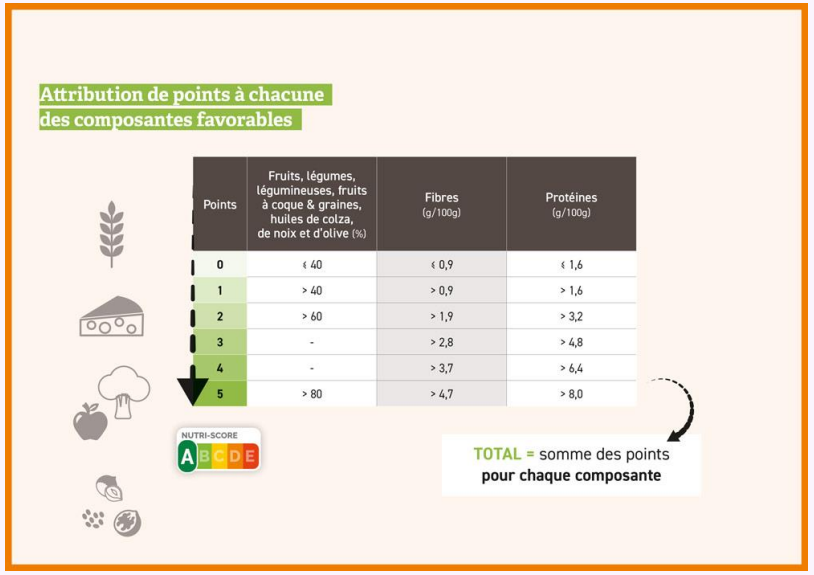


Le score est calculé en fonction de la **composition d'un produit** (pour 100 g).
Un **nombre de points est attribué à chaque composante**, en fonction de sa présence dans l'aliment.

Attribution de points à chacune des composantes défavorables

Points	Énergie (kJ/100g)	Sucres (g/100g)	Graisses saturées (g/100g)	Sodium (mg/100g)
0	≤ 335	≤ 4,5	≤ 1	≤ 90
1	> 335	> 4,5	> 1	> 90
2	> 670	> 9	> 2	> 180
3	> 1005	> 13,5	> 3	> 270
4	> 1340	> 18	> 4	> 360
5	> 1675	> 22,5	> 5	> 450
6	> 2010	> 27	> 6	> 540
7	> 2345	> 31	> 7	> 630
8	> 2680	> 36	> 8	> 720
9	> 3015	> 40	> 9	> 810
10	> 3350	> 45	> 10	> 900

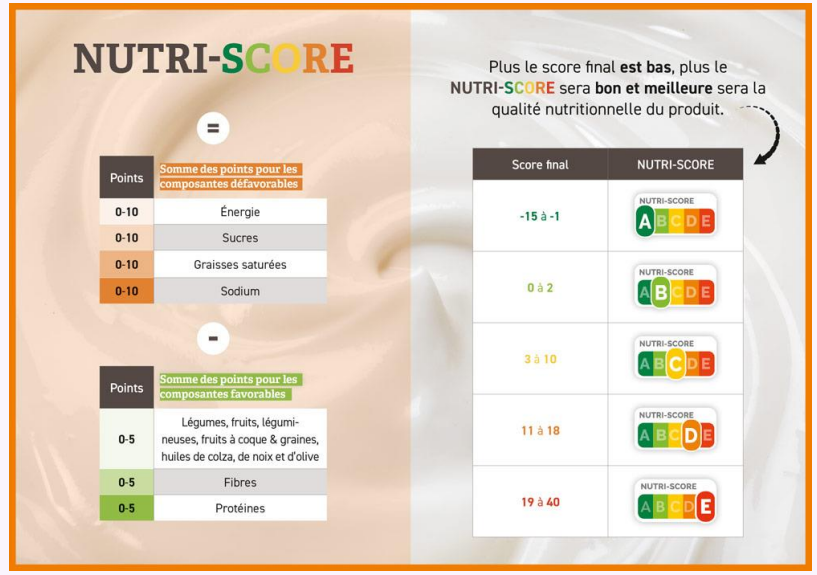
TOTAL = somme des points **pour chaque composante**



Attribution de points à chacune des composantes favorables

Points	Fruits, légumes, légumineuses, fruits à coque & graines, huiles de colza, de noix et d'olive (%)	Fibres (g/100g)	Protéines (g/100g)
0	≤ 40	≤ 0,9	≤ 1,6
1	> 40	> 0,9	> 1,6
2	> 60	> 1,9	> 3,2
3	-	> 2,8	> 4,8
4	-	> 3,7	> 6,4
5	> 80	> 4,7	> 8,0

TOTAL = somme des points **pour chaque composante**



Calcul des points négatifs

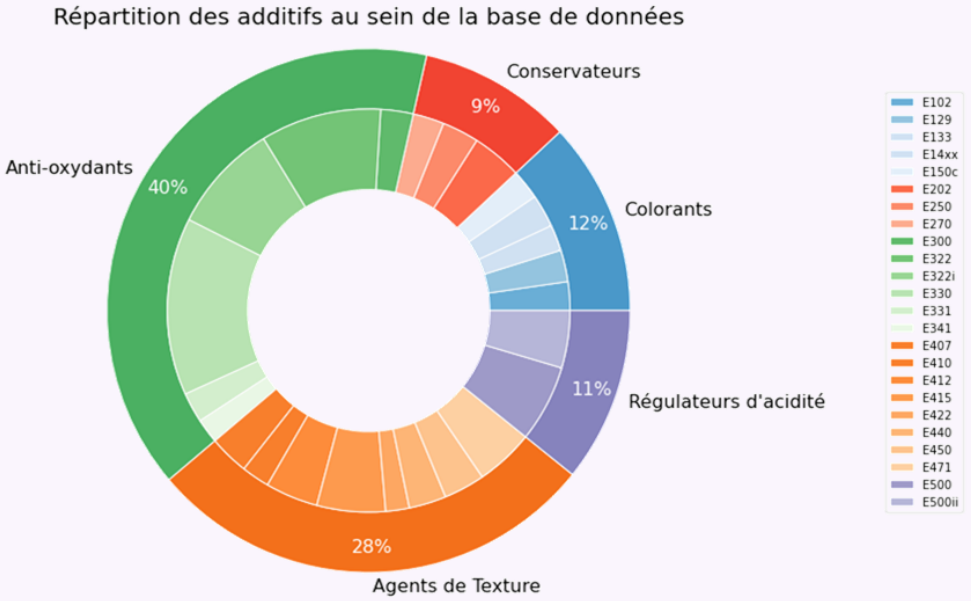
Création d'une colonne « points fruits et légumes » en fonction de la catégorie du produit

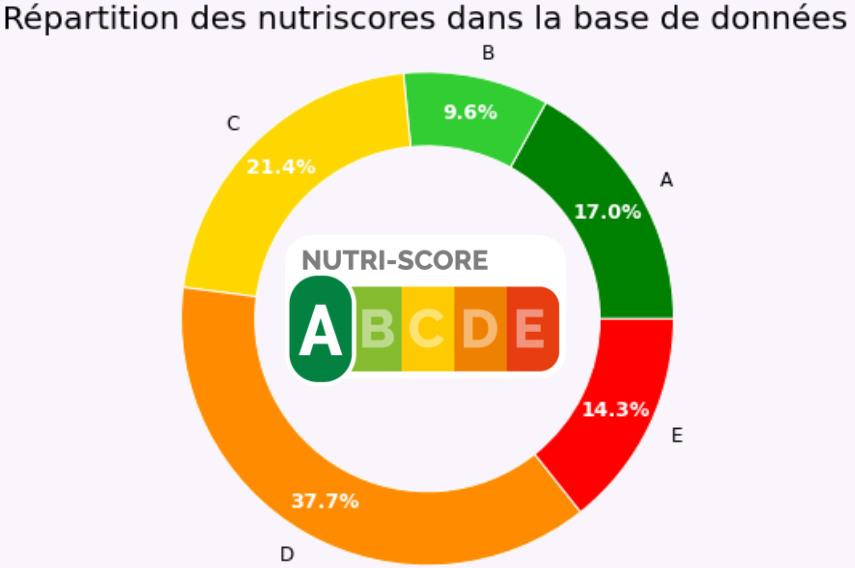
Attribution de la lettre

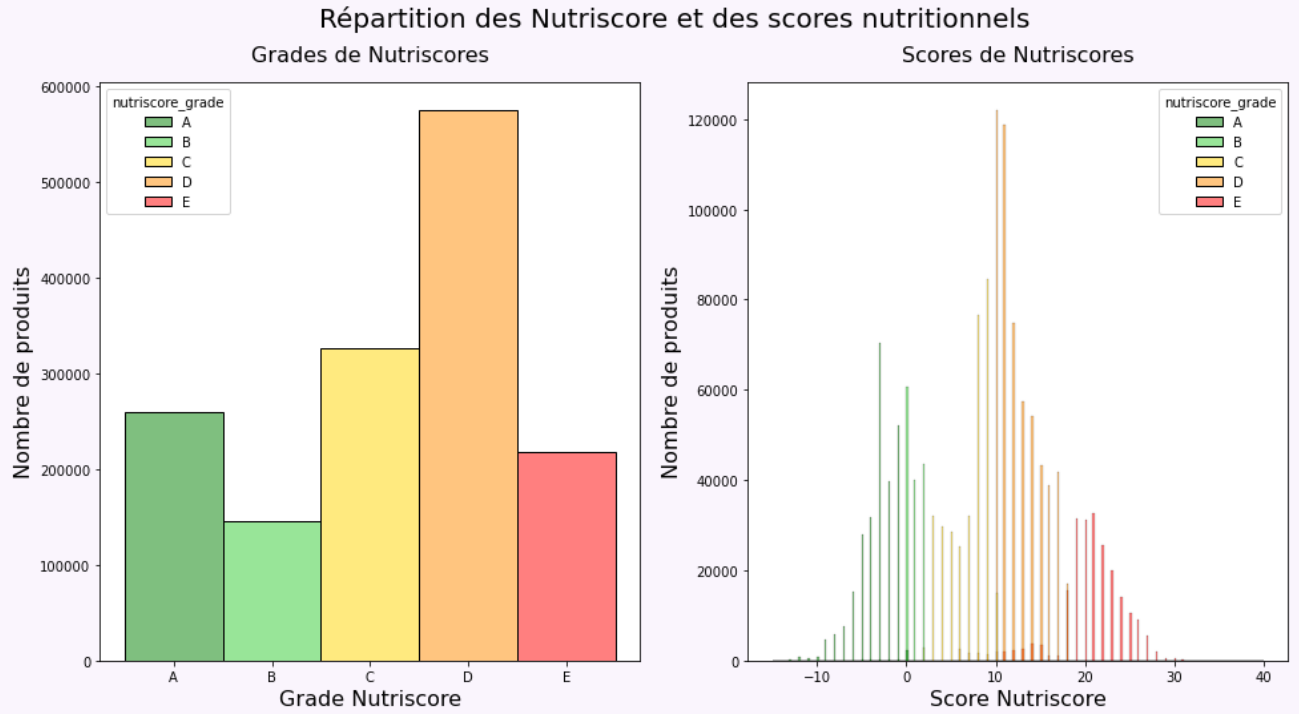
Calcul des points positifs

Calcul du score

Visualisation des données

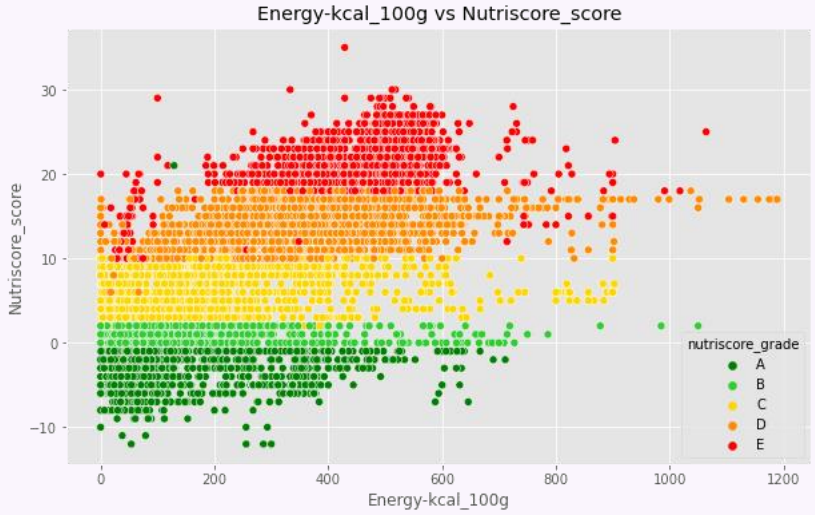


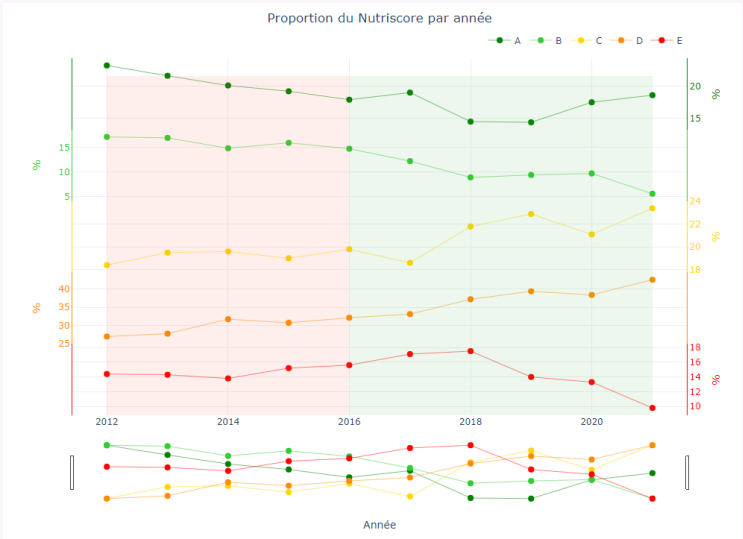


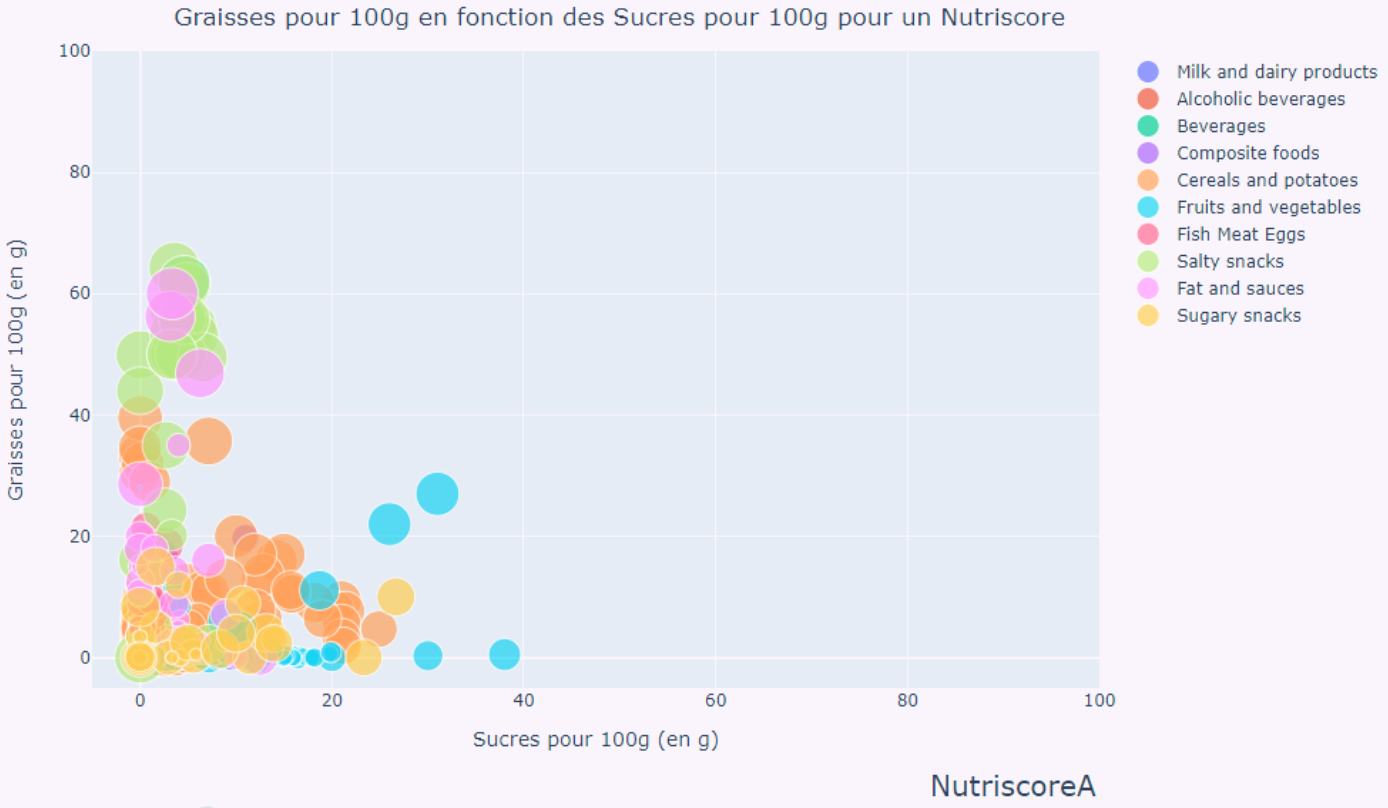


Visualisation des données

Visualisation interactive (1 variable qualitative vs 2 variables quantitatives)

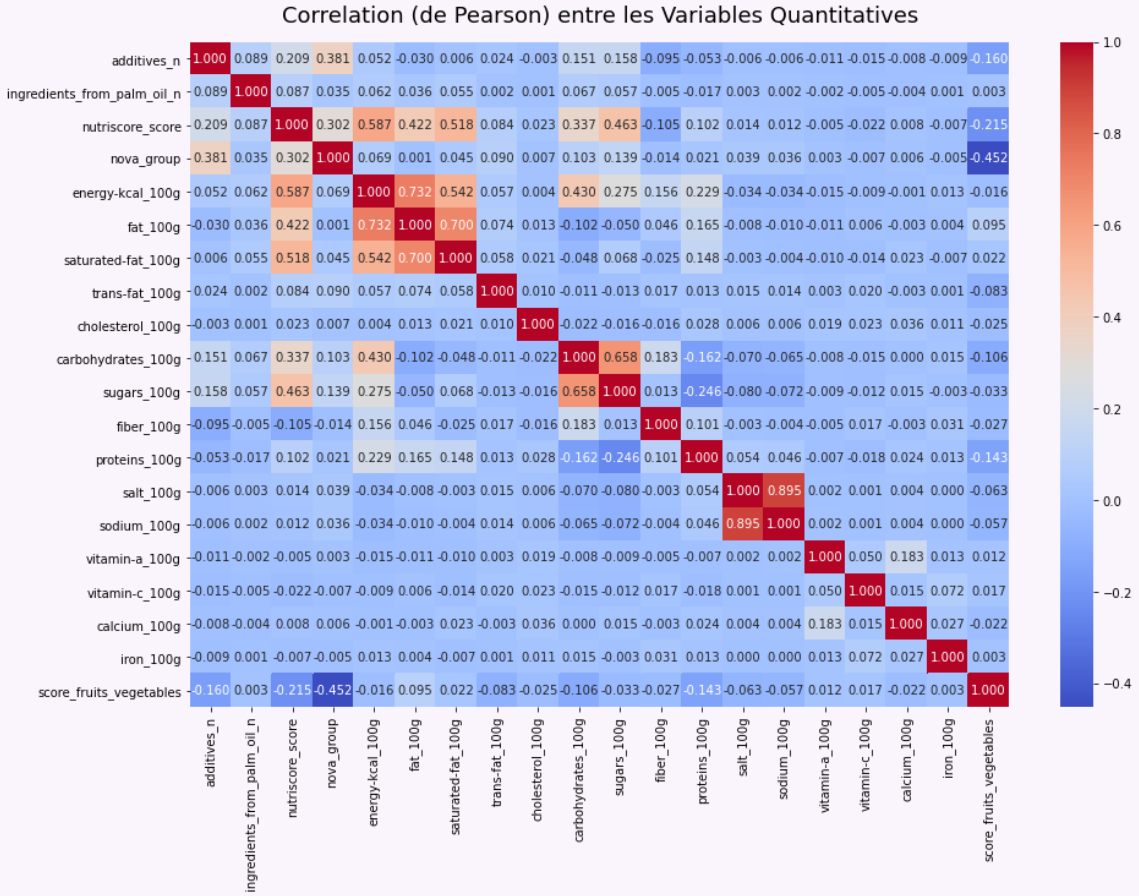




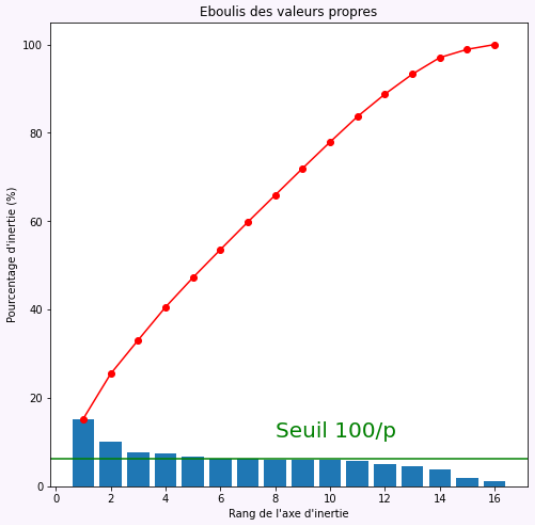


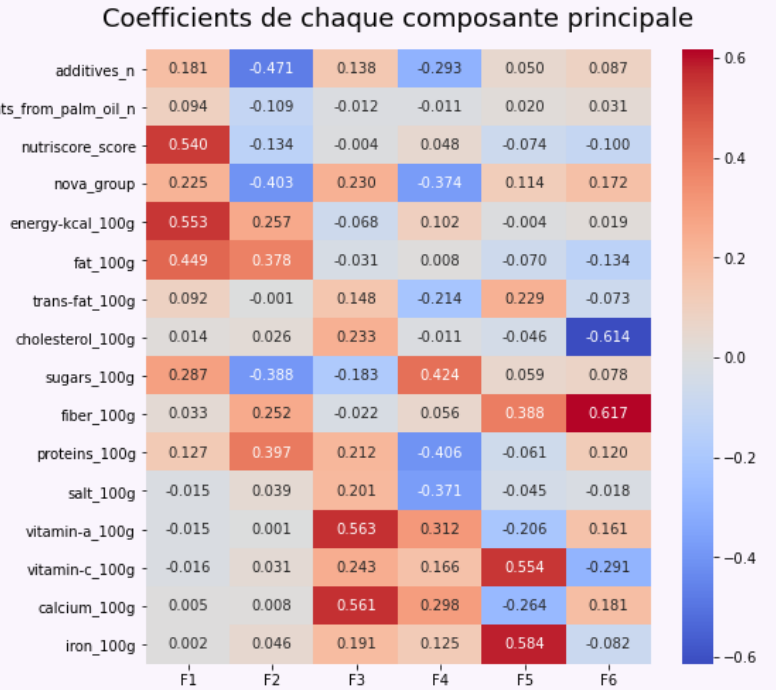
Analyse descriptive des données

Corrélation linéaire de Pearson



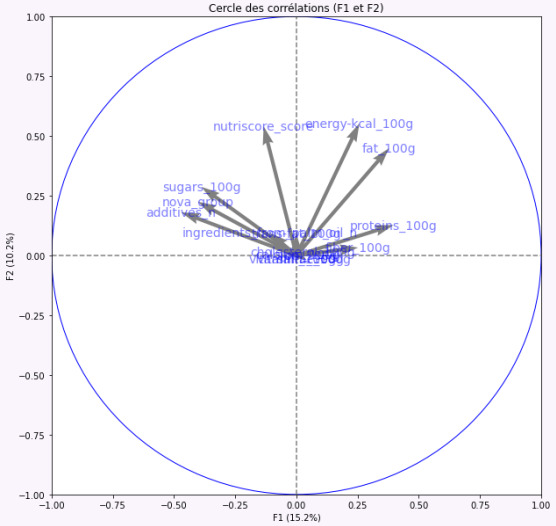
Analyse en composantes principales (ACP)

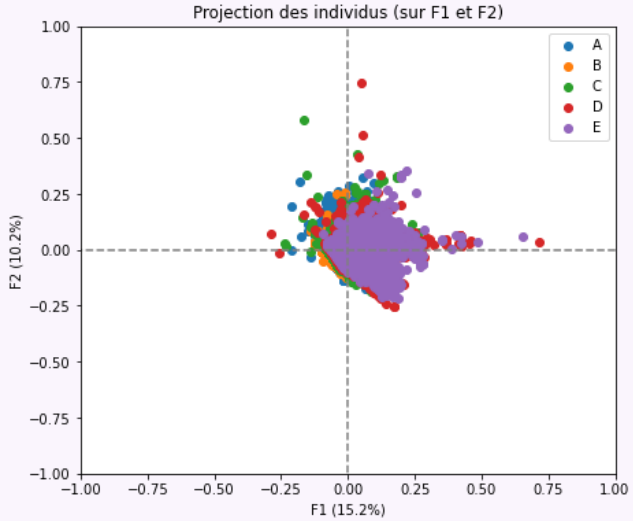




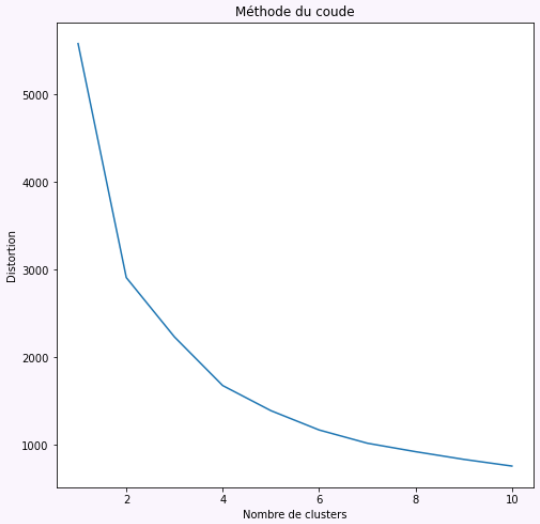
Analyse descriptive des données

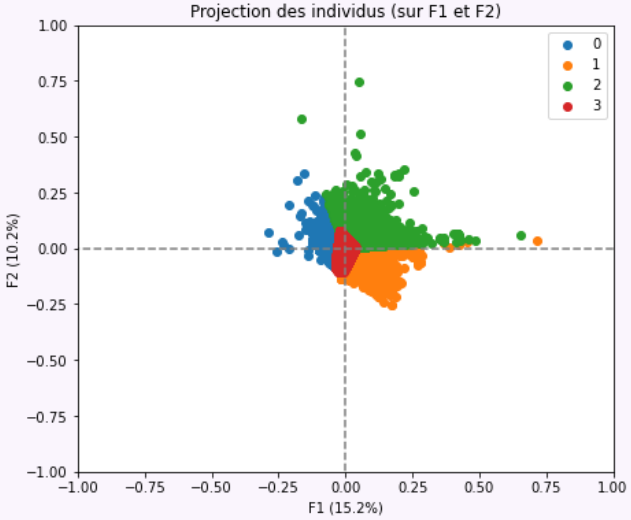
Analyse en composantes principales (ACP)





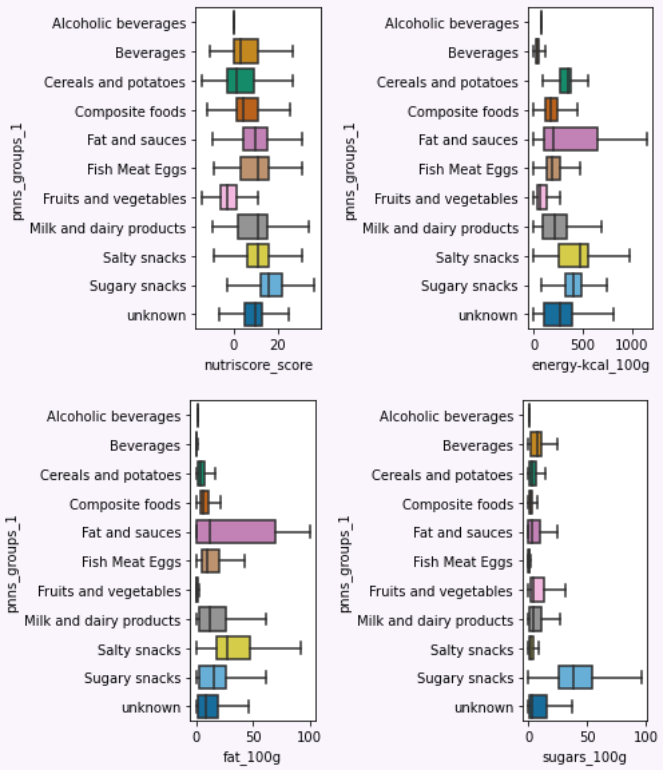
Partitionnement avec l'algorithme K-means





Analyse explicative des données

ANOVA



rapports de corrélation
(Variation interclasses / Variation totale)

	nutriscore_score	energy-kcal_100g	fat_100g	sugars_100g
labels_tags	0.008	0.002	0.001	0.001
nutriscore_grade	0.909	0.325	0.153	0.239
pnns_groups_1	0.199	0.165	0.132	0.240
countries	0.004	0.002	0.002	0.009

Test du CHI-2

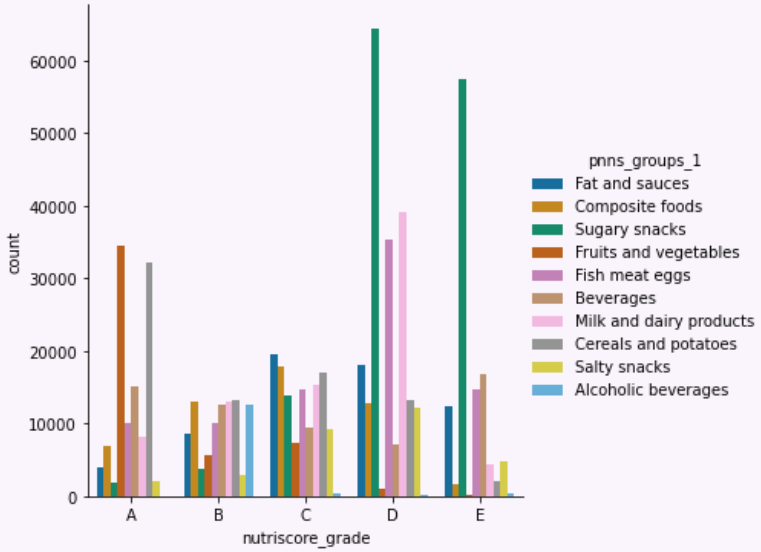
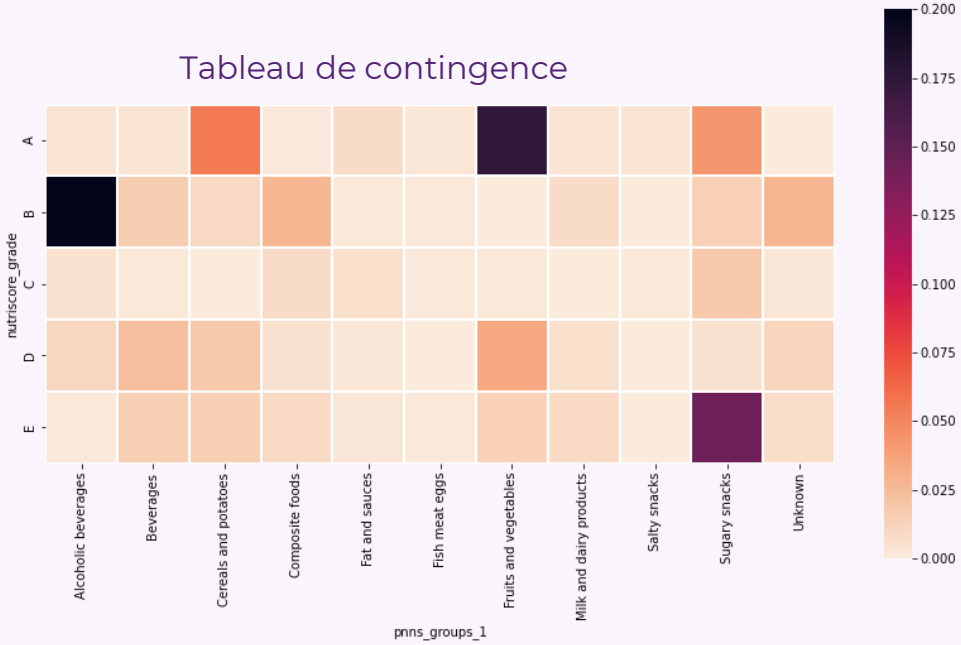


Tableau de contingence



Compétences

Nettoyage des données : sélection des variables et imputation des valeurs manquantes

Visualisation de chaque variable avec la représentation adéquate suivant leur distribution

Visualisation multivariée grâce aux visualisations interactives

Corrélation linéaire entre les variables quantitatives

Analyse descriptive : Analyse en Composantes Principales et Détermination de clusters : algorithme K-means

Analyse explicative : ANOVA / CHI-2