Contexte du projet

Nettoyage des données

Supprimer :

- valeurs aberrantes
- colonnes redondantes
- colonnes inutiles

Valeurs aberrantes

Valeurs dupliquées dans colonnes:

- code : 111
- product_name : 81663
- generic_name : 14211

Valeurs aberrantes

Action sur groupes de colonnes:

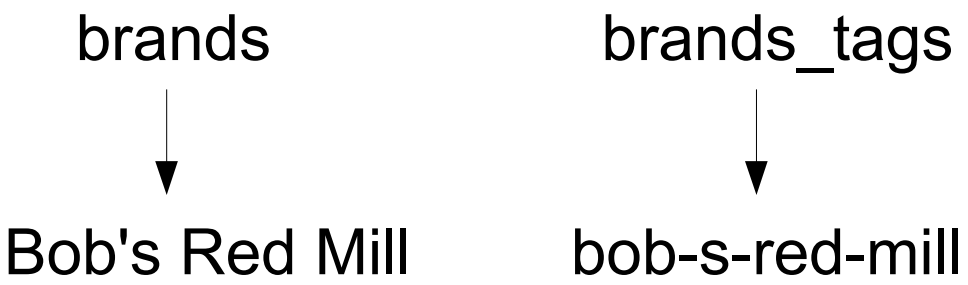
- valeurs négatives
- somme colonnes <= valeur spécifique
- somme colonnes <= colonne spécifique

Colonnes redondantes

- Marque
- Additifs
- Pays
- Nutri-grade
- Sel

Colonnes redondantes

Marque :



Colonnes redondantes

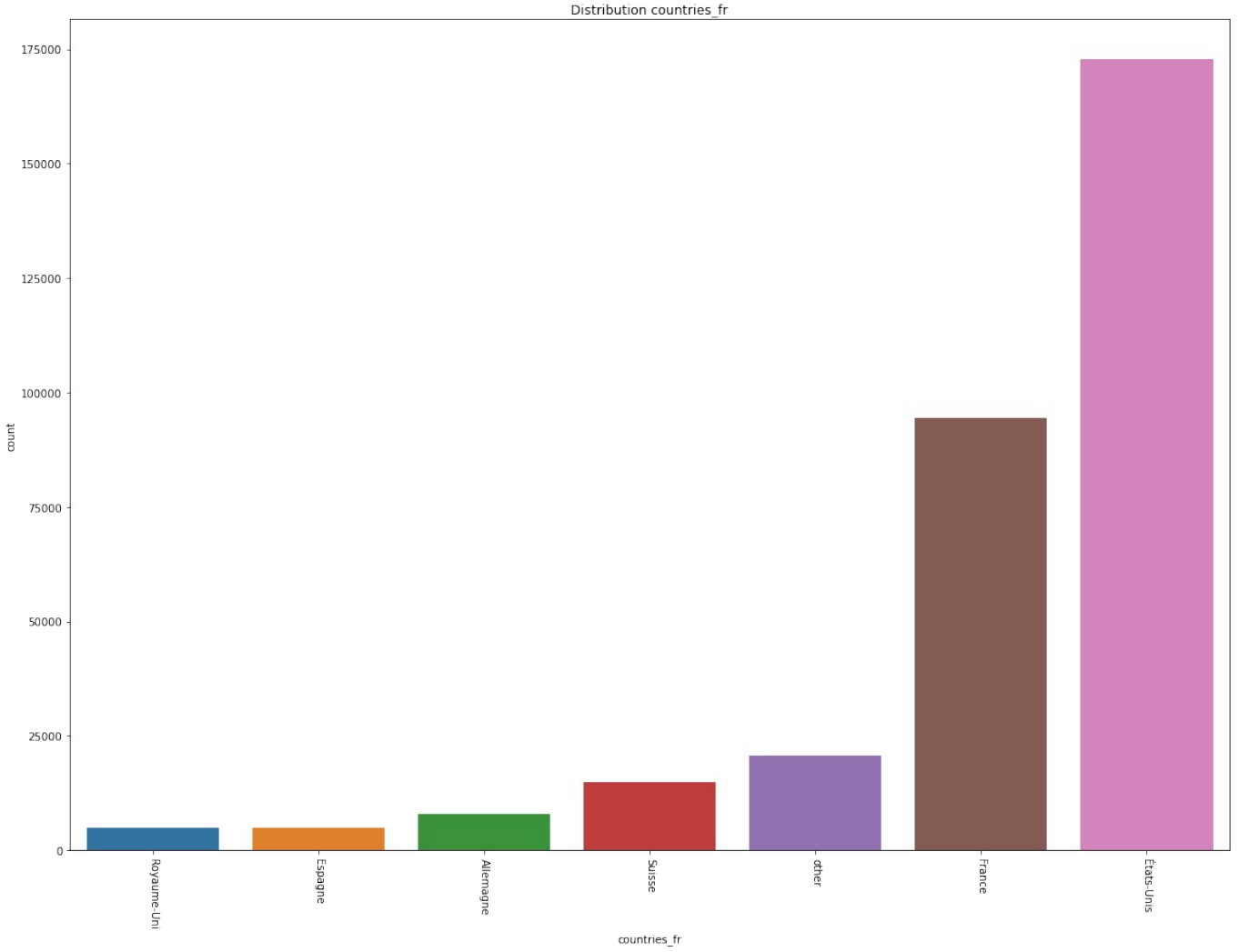
Additifs :

additives_tags → en:e100

additives_fr → E100 - Curcumine

Colonnes redondantes

Pays :



Colonnes redondantes

Nutri-score :

nutri-score_uk → formule originale

nutri-score_fr → formule modifiée

Colonnes redondantes

Sel :

salt = 2,5 X sodium

Analyse des données

Imputation des données

1) sélection des colonnes à imputer

2) Imputation par :

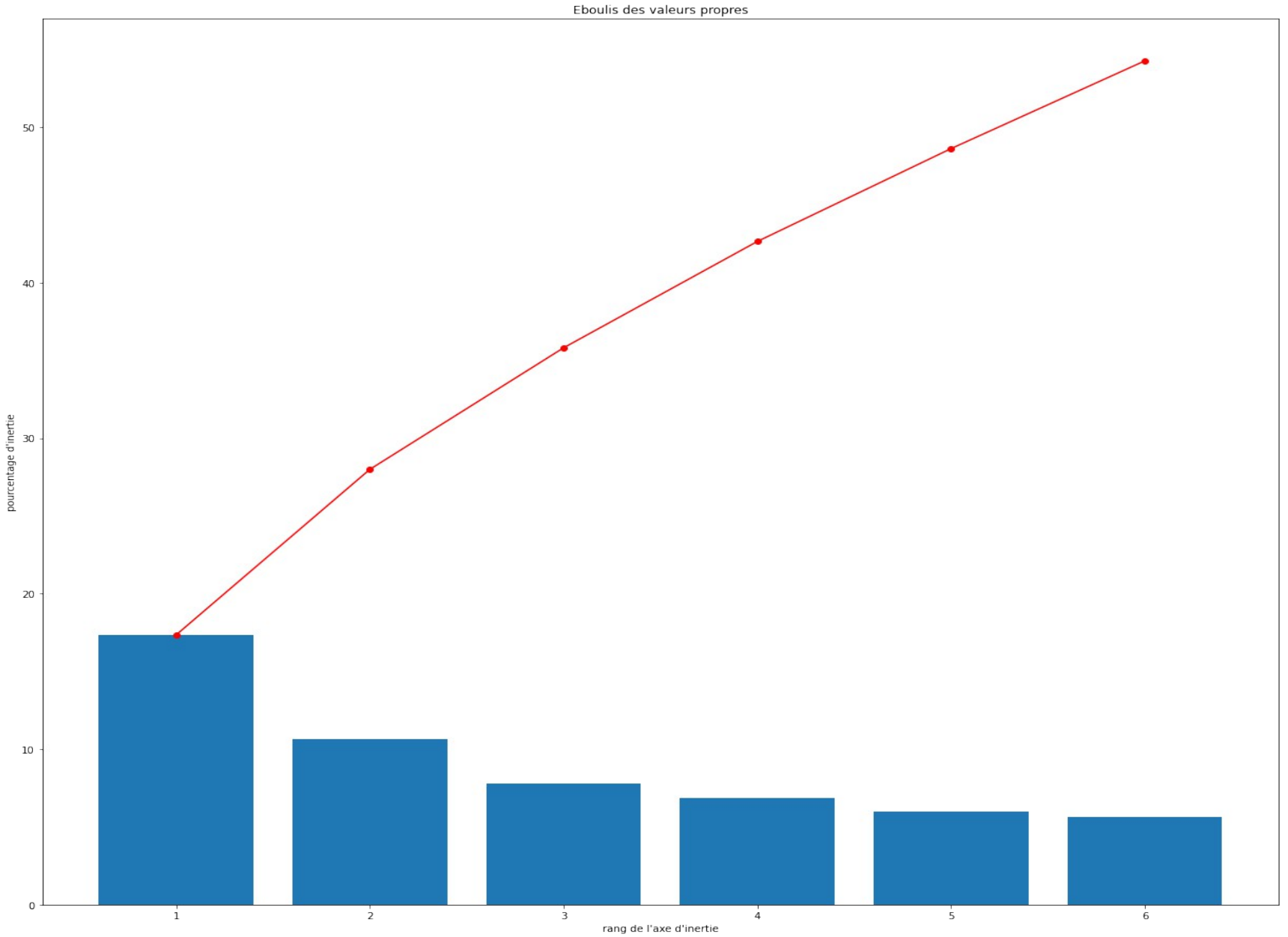
- zéro
- médianes des groupes pnns
- KNN
- médiane

Imputation des données

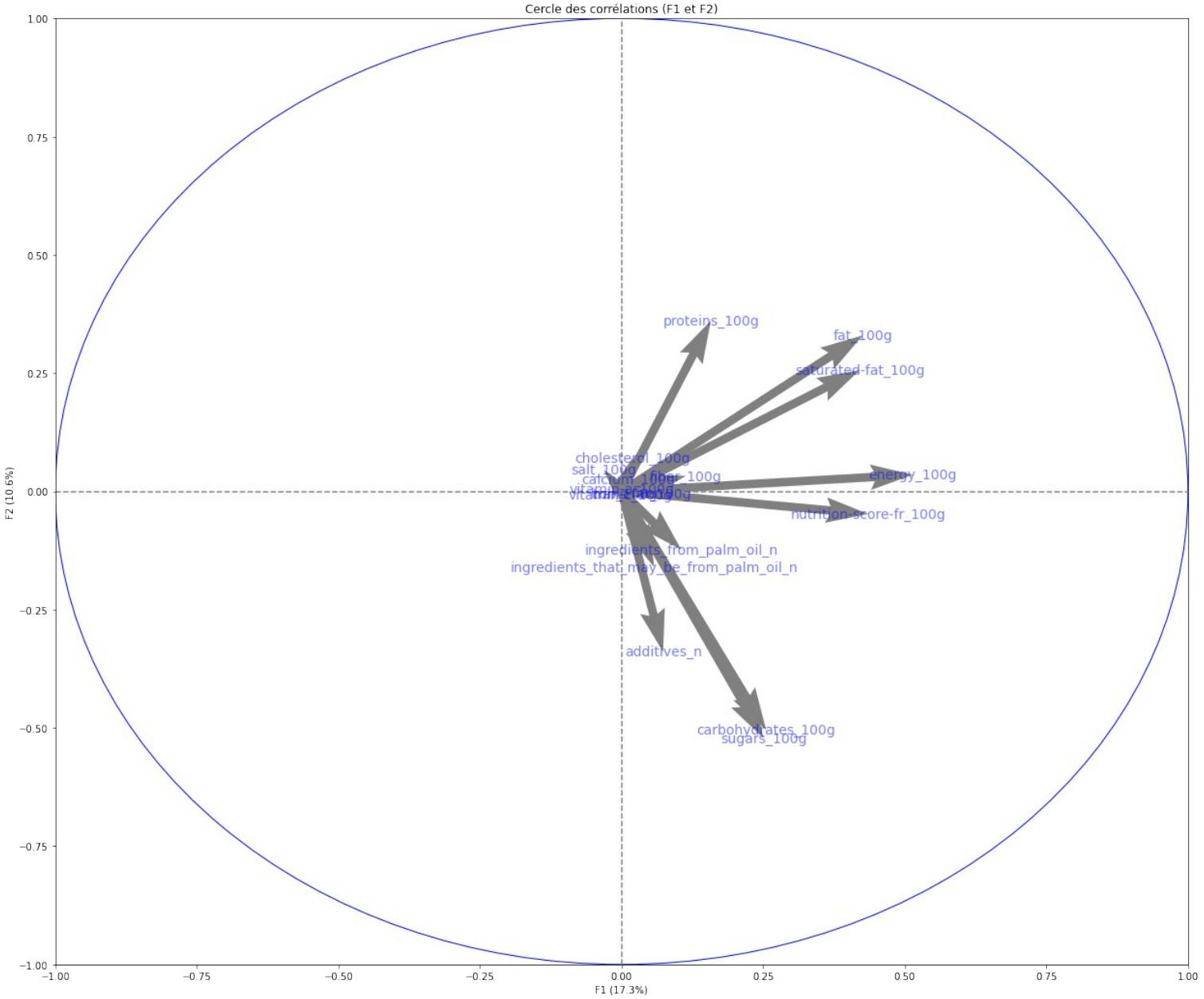
Imputation avec les groupes pnns

pnns_groups_1	pnns_groups_2
Beverages	Artificially sweetened beverages
	Fruit juices
	Fruit nectars
	Non-sugared beverages
	Sweetened beverages

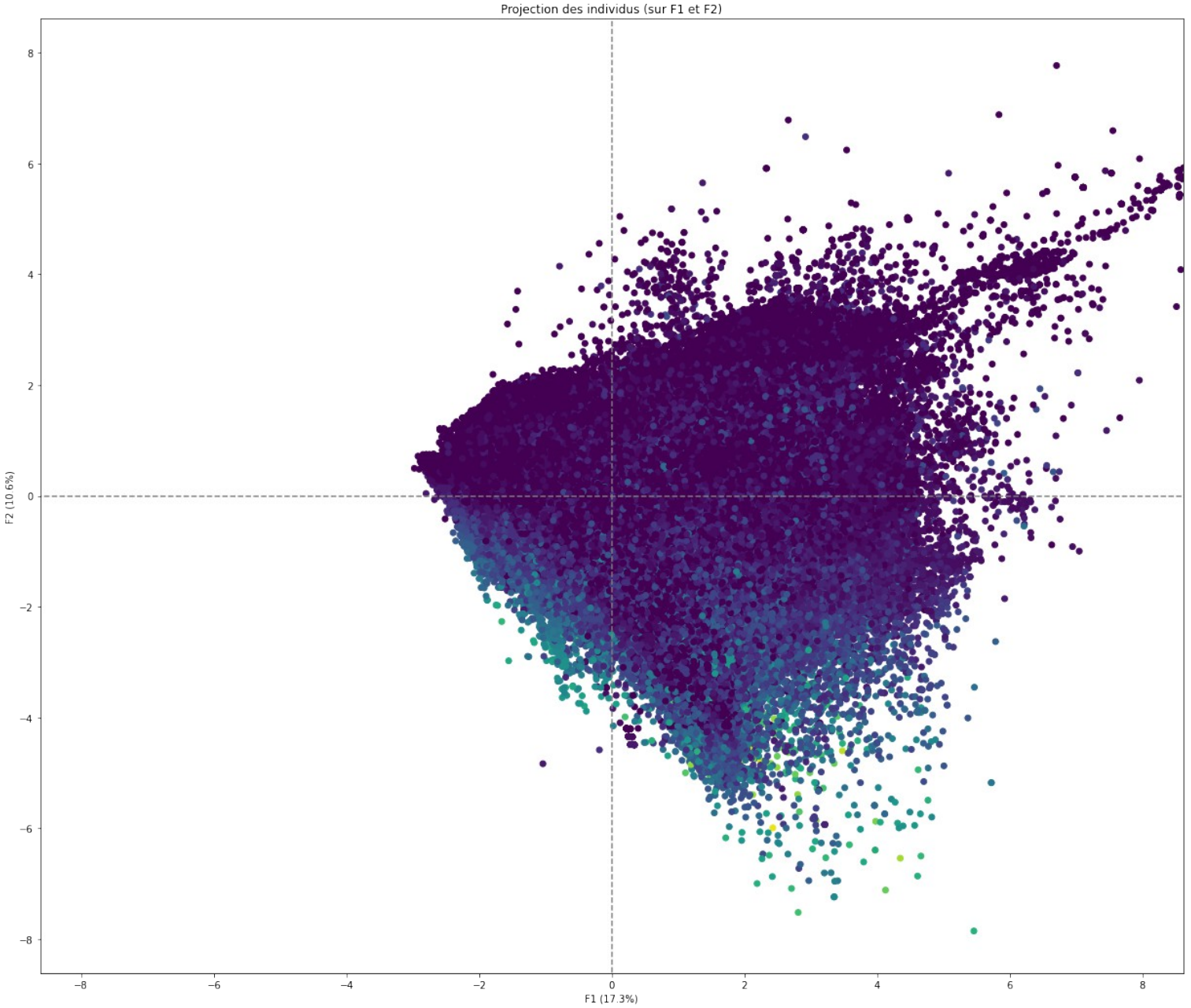
Analyse en composantes principales



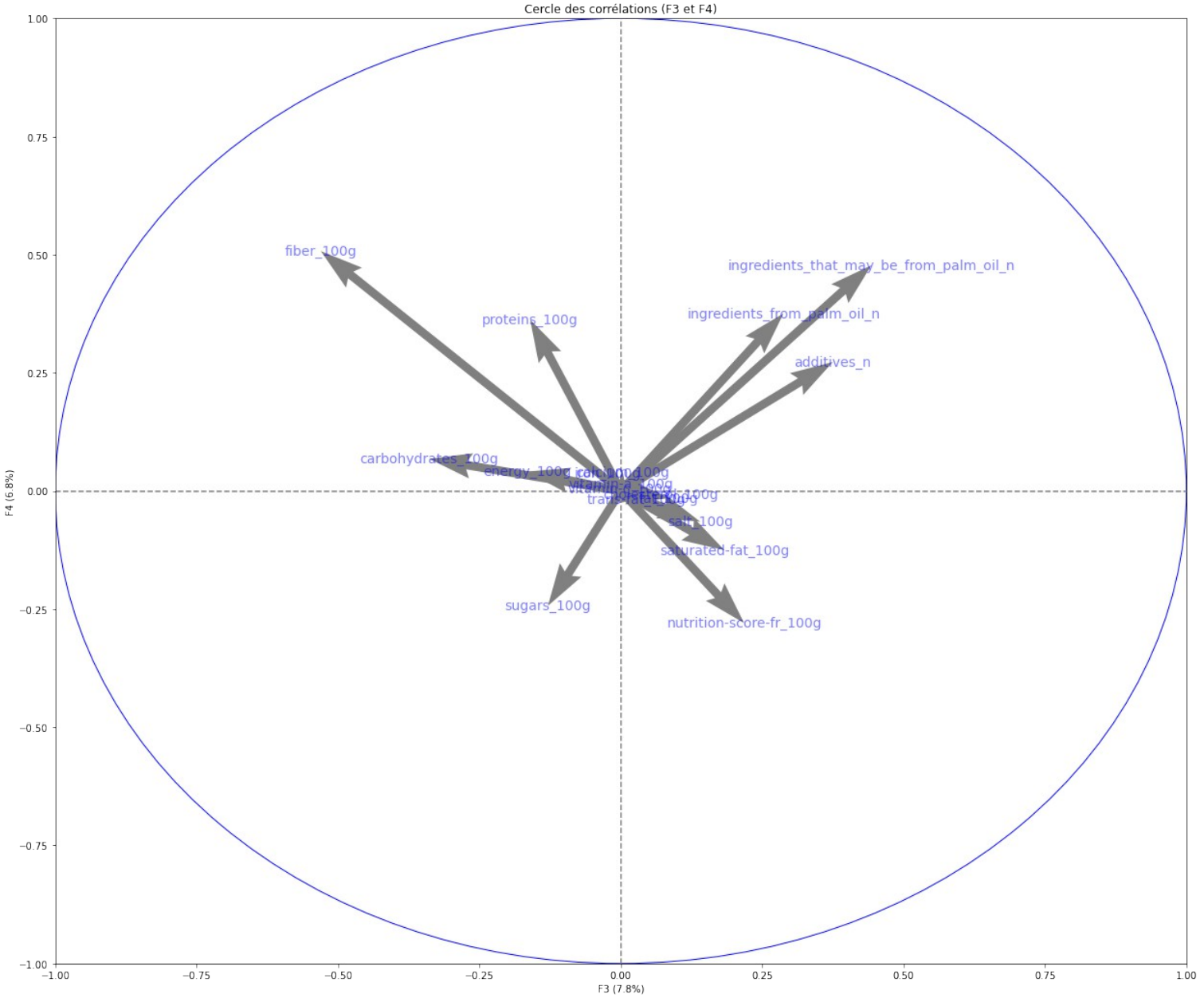
Analyse en composantes principales



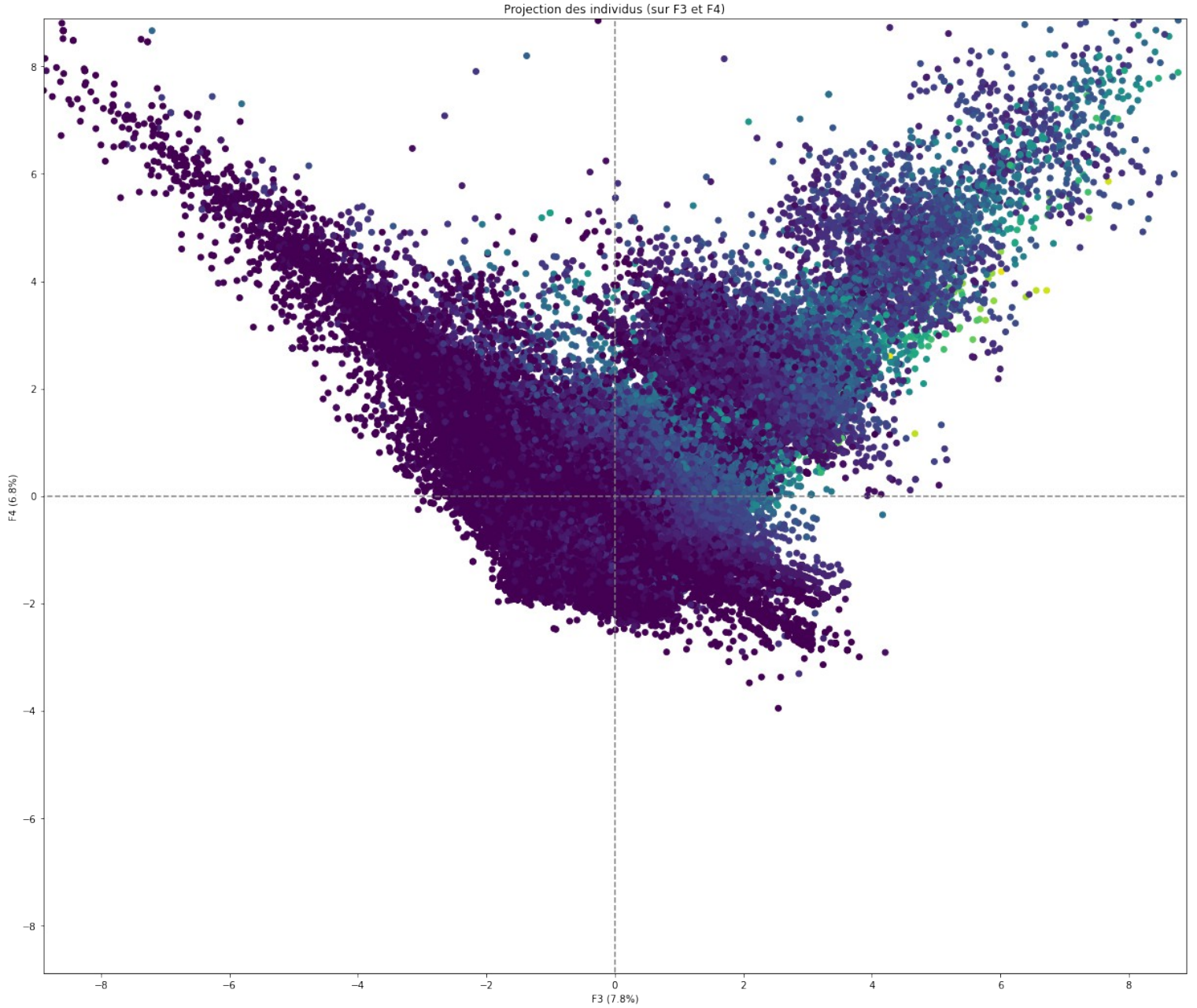
Analyse en composantes principales



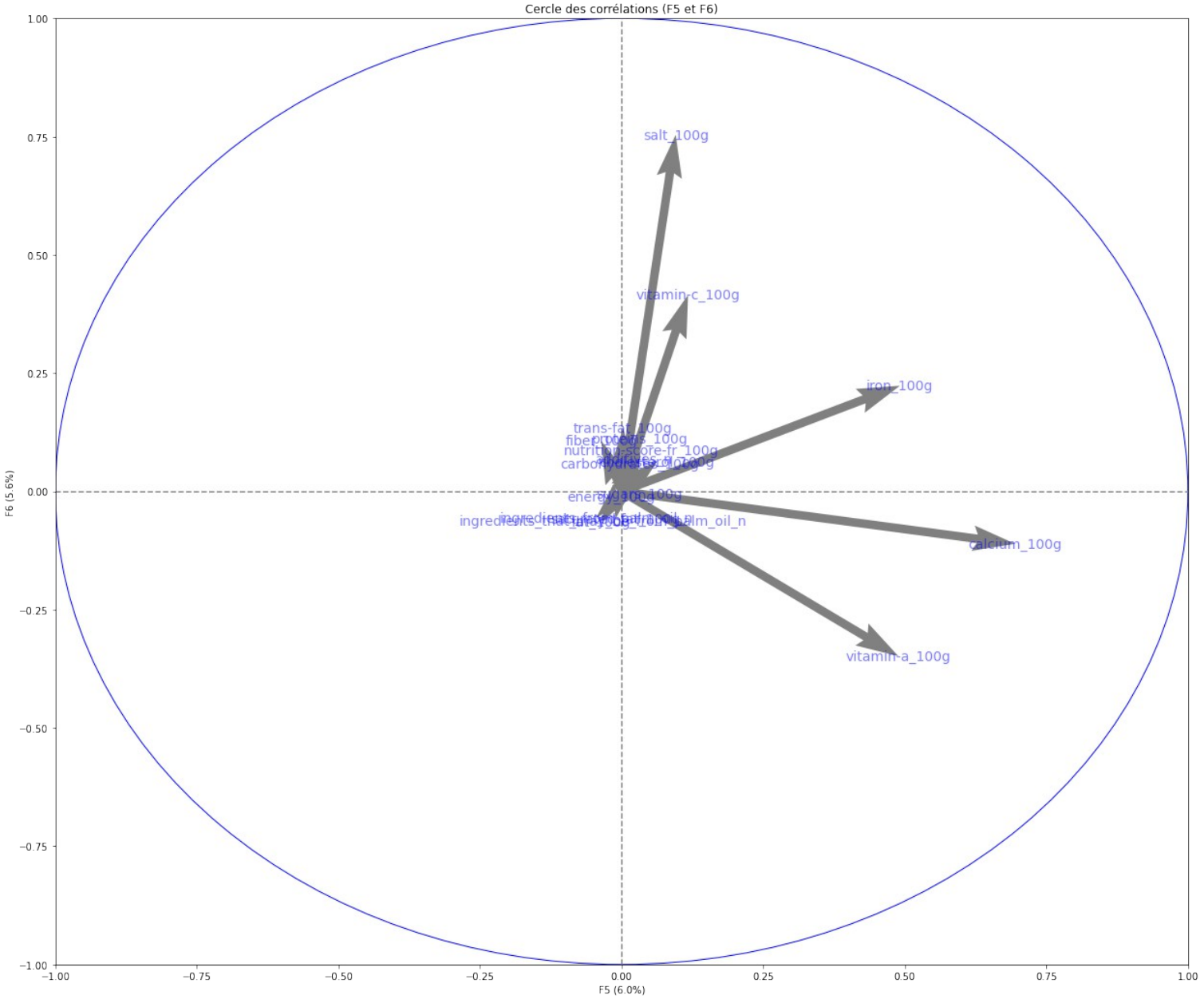
Analyse en composantes principales



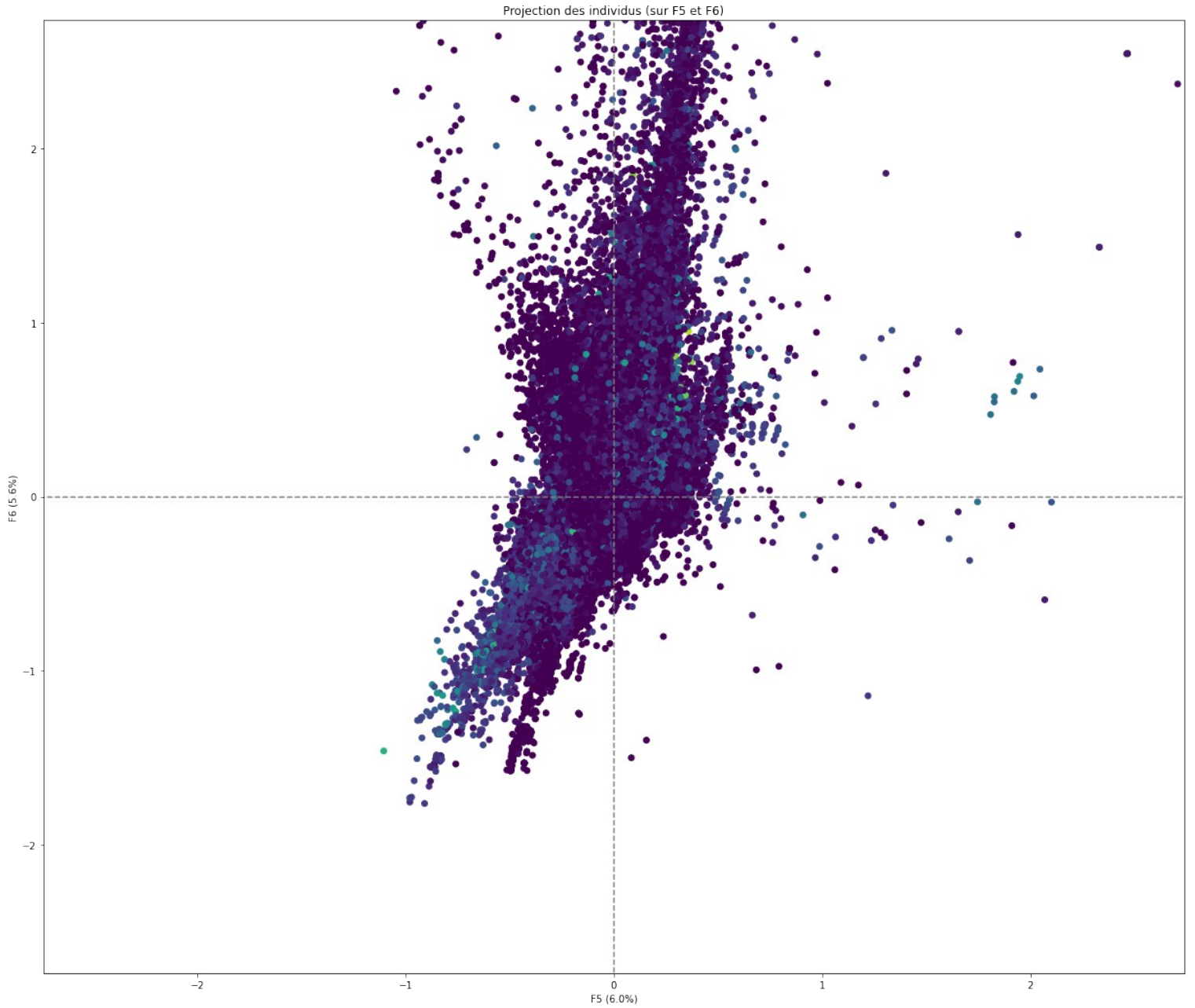
Analyse en composantes principales



Analyse en composantes principales



Analyse en composantes principales



Application

- centrée sur les additifs

- recommandation de produits

- système de réputation des marques

Axes d'améliorations

- plus d'information sur les additifs

- imputation avec groupes pnns