

Parcours Data
scientist

Projet 3 :
Concevez une
application au
service de la
santé publique



Sommaire

- 1 Idée d'application
- 2 Opérations de nettoyage effectuées
- 3 Description et analyse univariées et bivariées
- 4 Analyse multivariée et résultat statistiques associés
- 5 3 observation sur la pertinence et la faisabilité du projet

Synthèse



1

Idée d'application



Ingrédient 1 :

tomate

Ingrédient 2 :

poivron

Ingrédient 3 :

aubergine

Ingrédient 4 :

Ingrédient 5 :

Rechercher

The image displays three jars of tomato sauce, each with its Nutri-Score label. The first jar is 'Sauce tomate BIO AUBERGINES & POIVRONS' with a Nutri-Score of 'A'. The second jar is 'Tomates poivrons aubergines' with a Nutri-Score of 'A'. The third jar is 'MOUTARDE' with a Nutri-Score of 'A'. Each label features the Nutri-Score logo with the letter 'A' in a green circle, followed by 'B' (yellow), 'C' (orange), 'D' (red), and 'E' (dark red).

L'idée

fournir à
l'utilisateur une
application qui
permettent de
trouver les produits
ayant le meilleur
nutriscore à partir
de mots clefs.

L'exemple

si l'utilisateur tape :
« tomate »,
« poivron » et
« aubergine », voici
les trois premiers
résultats qu'il verra
apparaître.



2

Opérations de nettoyage effectuées



Phase 1 :
Taille 1,890,337 x 186
80% de données manquantes

Sélection des produits ayant un nom et étant **vendu en France**.

Suppression des colonnes ayant moins de **5% de valeurs** renseignées

Phase 2 :
800,080 x 76
49% de données manquantes

Sélection des colonnes nécessaire pour :

- **nutriscore**

energy-kj_100g, saturated-fat_100g, fat_100g,
carbohydrates_100g, sugars_100g, fiber_100g,
proteins_100g, salt_100g
nutriscore-score-fr-100g, nutriscore-grade

- **recherche et classement:**

product_name, brands, categories, pnns_groups_1,
pnns_groups_2, image_url



Suppression des lignes pour
lesquelles **aucunes des**
données nutritionnelles
sélectionnées ne sont indiquées

Suppression des **doublons**

Phase 3 :
627,333 x 20
22% de données
manquantes

Les données aberrantes



Nutritif supérieur
à 100g pour 100g
de produit.

Kilo joule > 3766

Fibres > 20

Protéines > 88

Les données manquantes

Par régression

prédiction du **nutriscore** en fonction des valeurs nutritionnelles

régression linéaire score **56%**

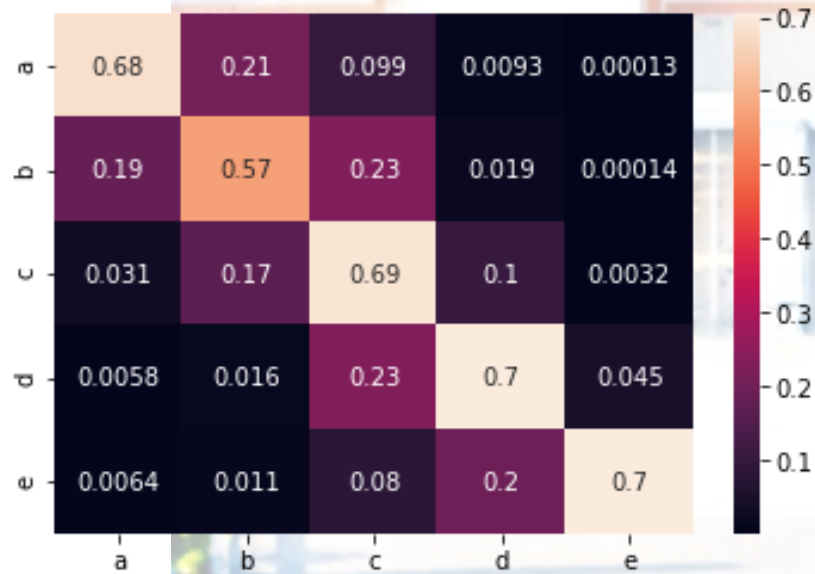
régression KNN score **84%**

Par classification (KNN)

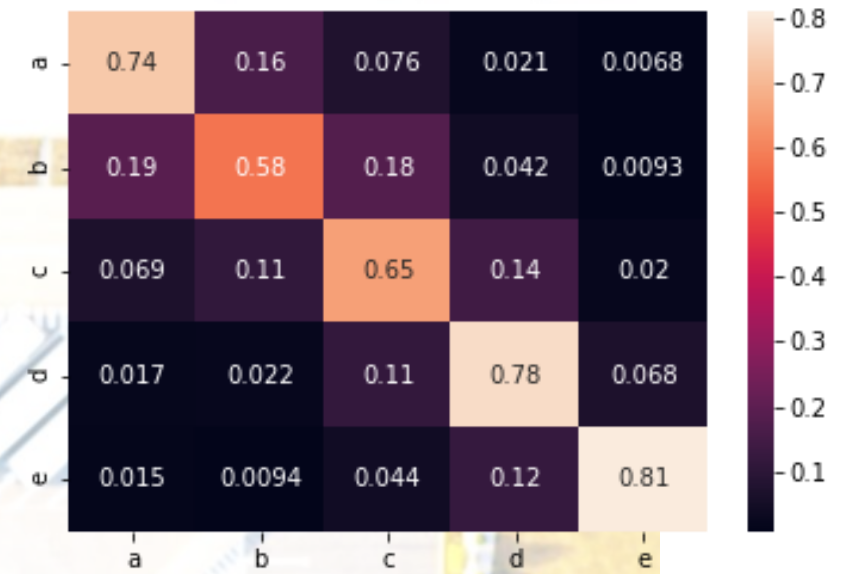
prédiction du **nutrigrade** en fonction des valeurs nutritionnelles

score **72%**

Régression



Classification



	energy-kj_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	y_pred_R	y_pred_C	y_test
0	2198.0700	27.0	18.2	62.9	29.8	6.7	0.20	e	e	e
1	1256.0400	15.3	9.9	35.4	17.5	4.7	0.40	d	d	d
2	2122.7076	25.8	14.9	60.6	38.9	6.5	0.46	e	e	e
3	136.0000	0.5	0.1	5.4	5.4	10.0	0.50	a	a	e
4	598.0000	5.6	2.0	0.5	0.5	23.0	1.80	d	d	c

79%

tx de corrélation régression/classification

Remplacement des données manquantes
nutriscore et nutrigrade par les
prédictions.

Jeu de donné définitif :

579,132 x 16

11% de valeurs nulles

(uniquement dans les colonnes brands, categories,
image_url et fibre)

```
df_def.columns
```

```
Index(['product_name', 'brands', 'categories', 'pnns_groups_1',  
      'pnns_groups_2', 'image_url', 'energy-kj_100g', 'fat_100g',  
      'saturated-fat_100g', 'carbohydrates_100g', 'sugars_100g', 'fiber_100g',  
      'proteins_100g', 'salt_100g', 'nutriscore', 'nutrigrade'],  
      dtype='object')
```


3

Description et analyse univariées et bivariées



Boîte à moustache

2 exemples extrêmes

Salt

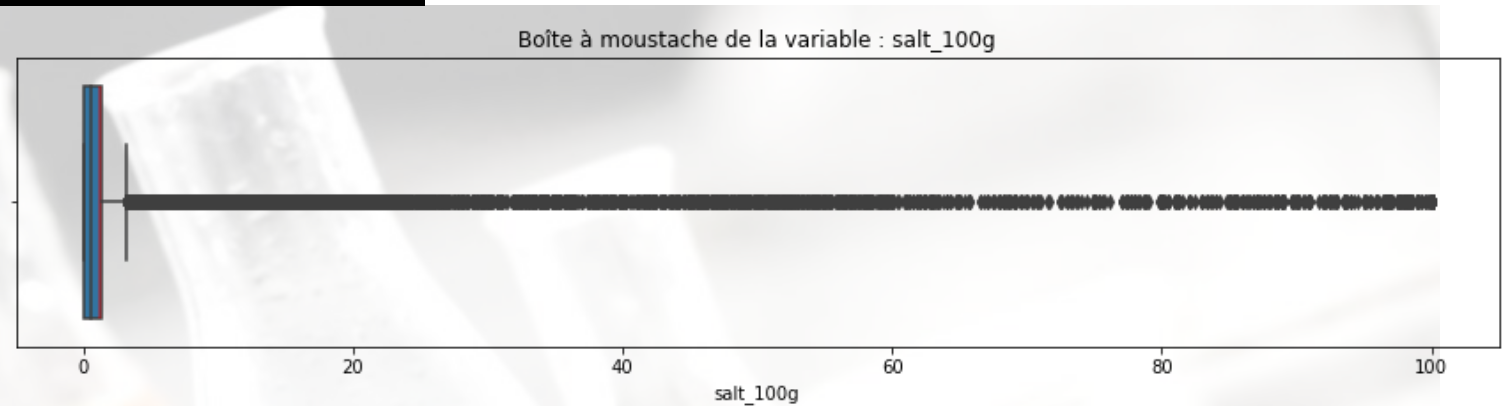
Min : 0

Max : 100

Moy : 1,26

Med : 0,56

Std : 4,2



Carbohydrates

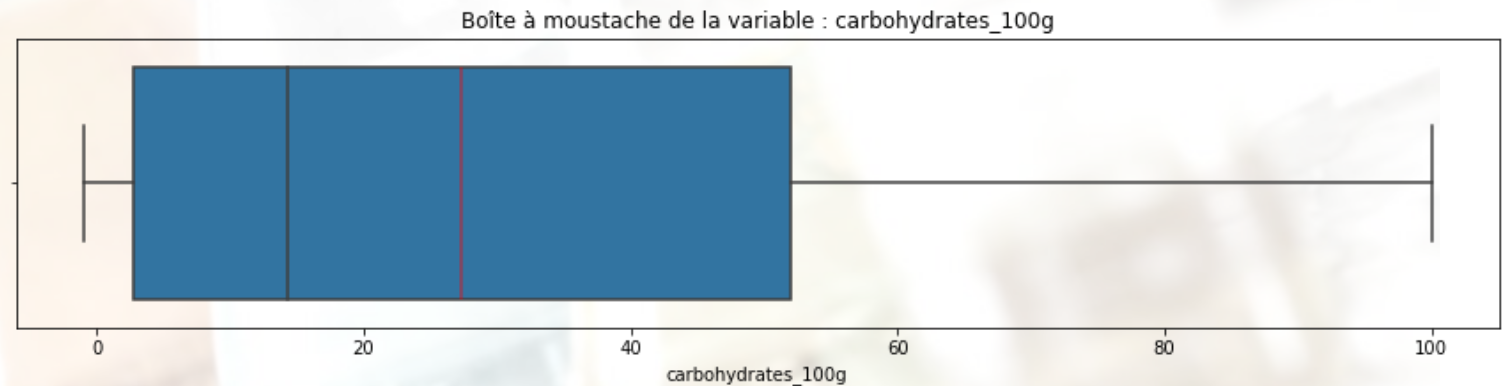
Min : 0

Max : 100

Moy : 27,34

Med : 14,3

Std : 27,60

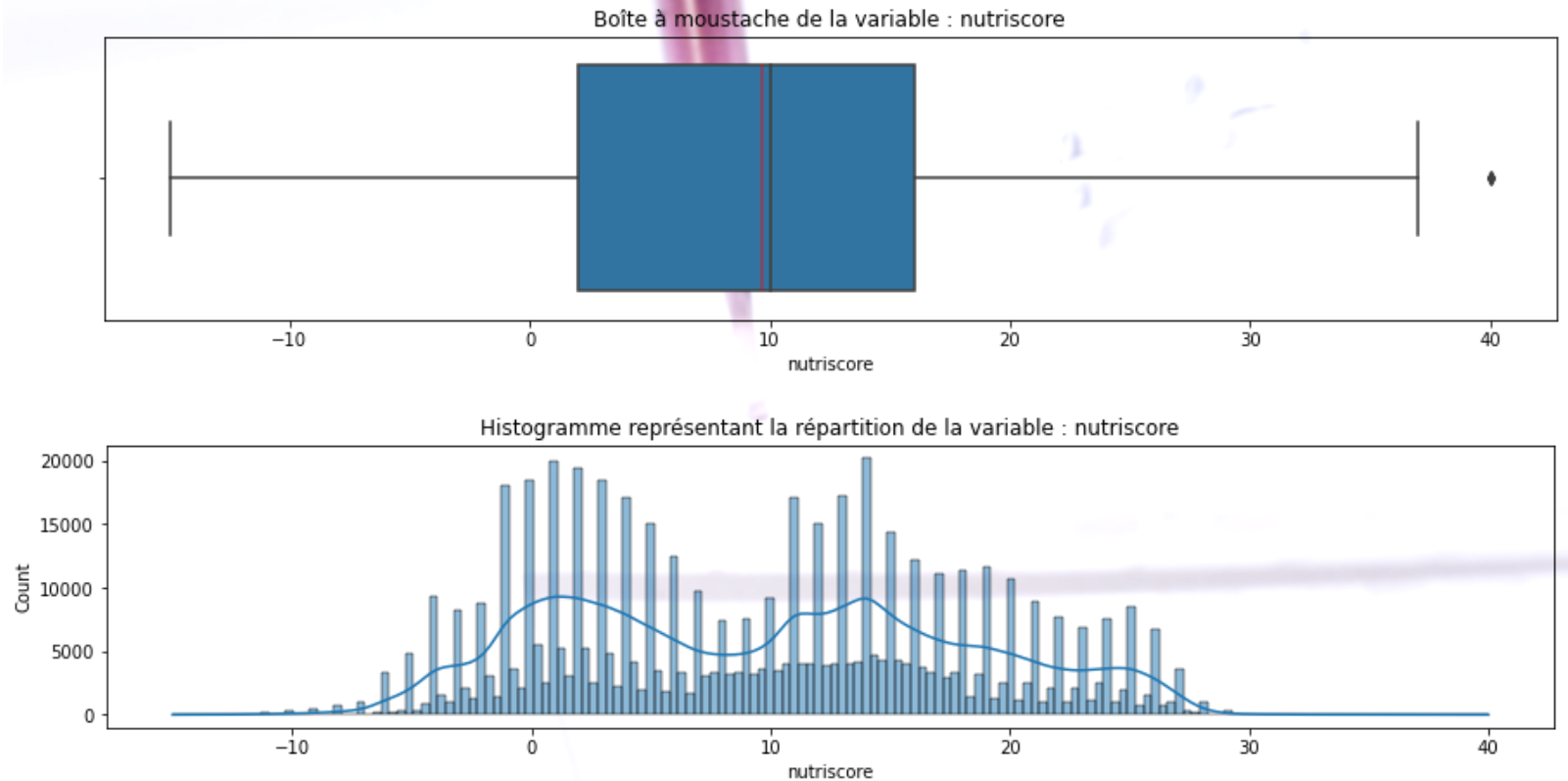


Variable nutriscore

Boîte à moustache et histogramme

Test de **normalité** sur
la variable
nutriscore.

Échec successif des
tests d'Anderson,
d'Agostino et de
Shapiro-Wilk.

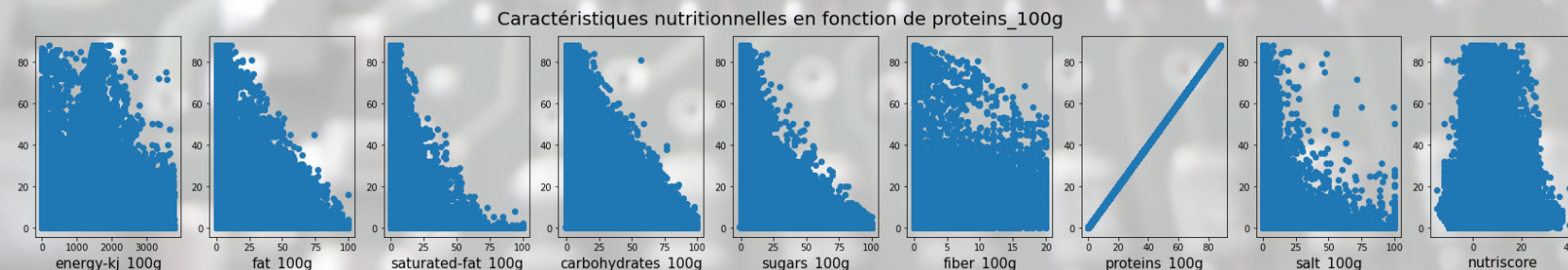
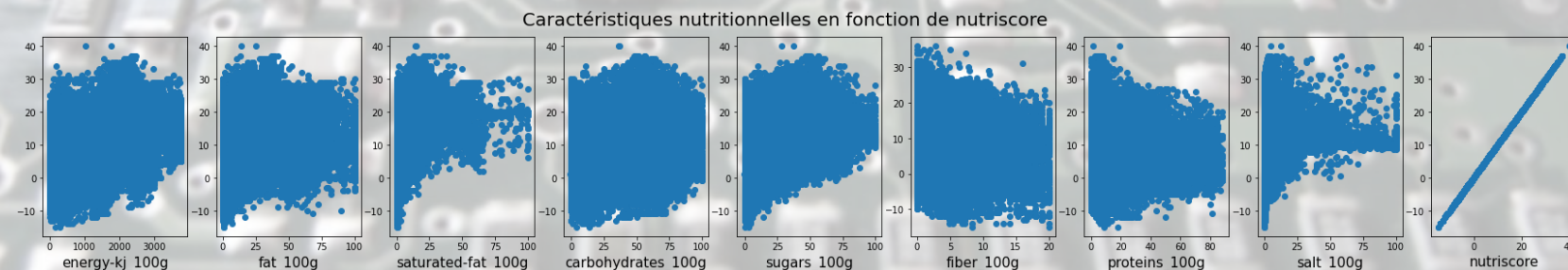
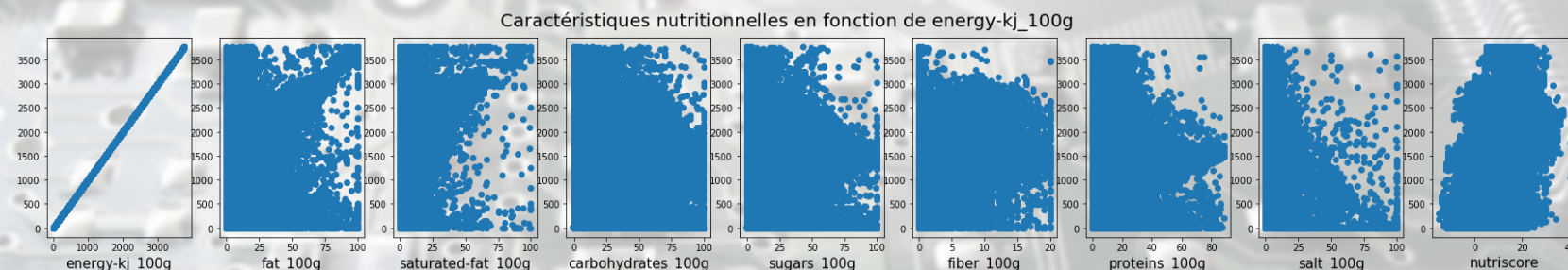


Analyses bivariées

Nuage de points entre toutes les variables quantitatives

Pas de **corrélation**
évidente entre les
variables mais pas de
franche
indépendance non
plus.

Coefficient de
Pearson de **0,61**
entre nutriscore et
energie

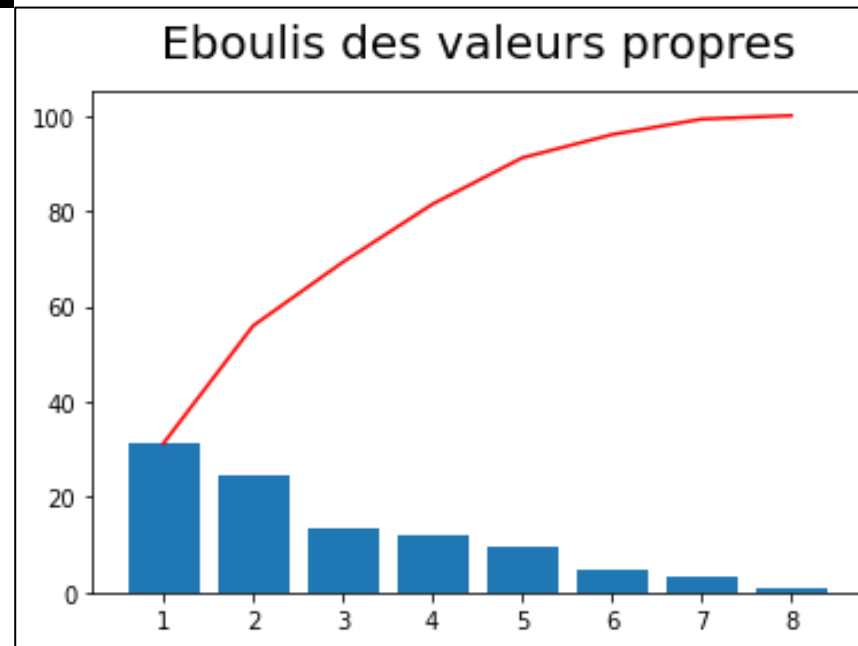
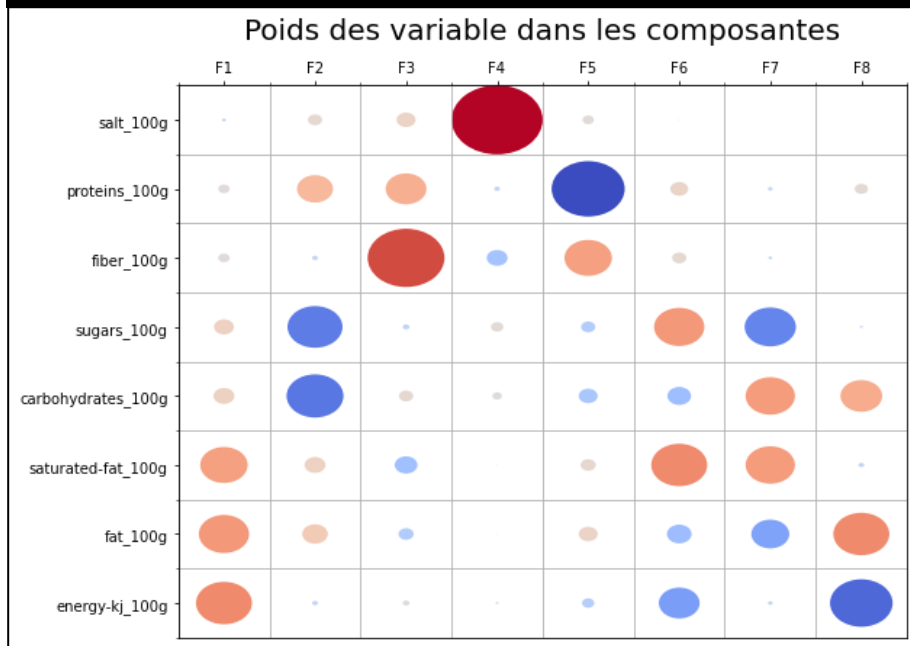


4

Analyse multivariée et résultat statistiques associées



ACP (analyse des composantes principales)

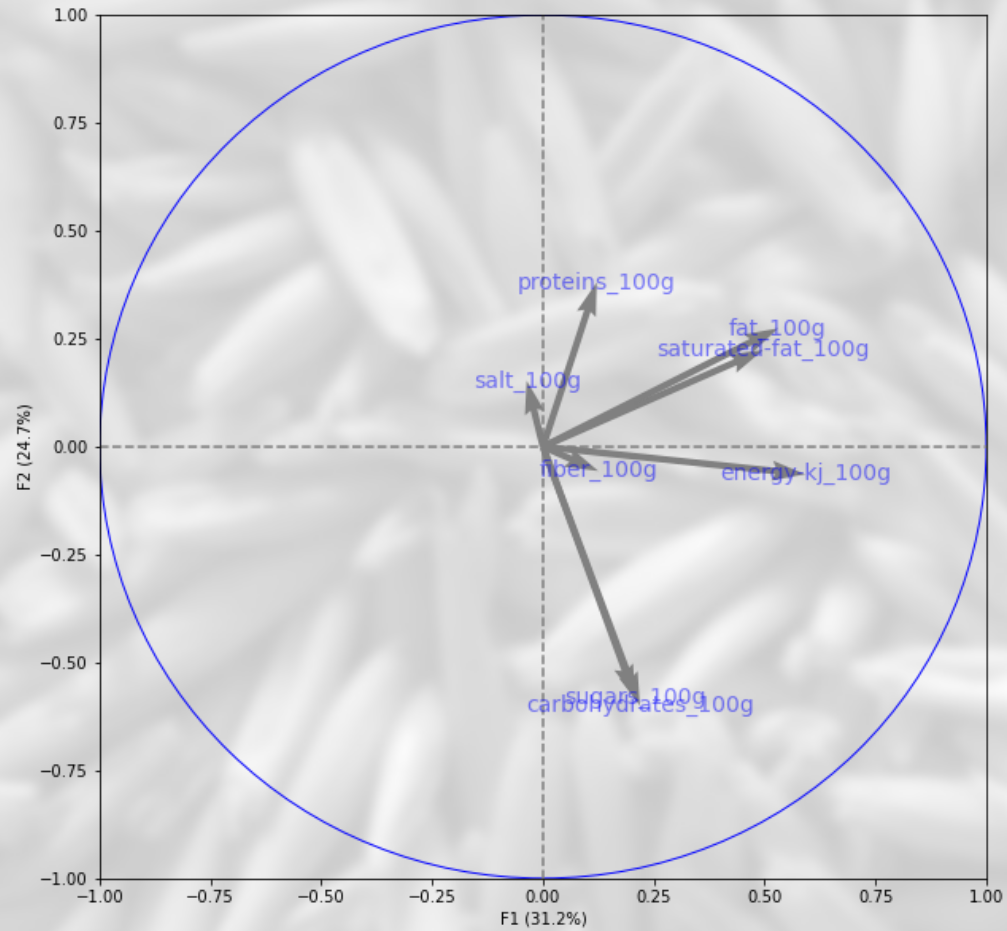


	F1	F2	F3	F4	F5	F6	F7	F8
energy-kj_100g	0.59	-0.06	0.07	0.03	-0.13	-0.43	-0.05	-0.66
fat_100g	0.53	0.27	-0.16	-0.01	0.20	-0.26	-0.40	0.59
saturated-fat_100g	0.50	0.22	-0.24	0.01	0.16	0.59	0.52	-0.06
carbohydrates_100g	0.22	-0.60	0.15	0.10	-0.20	-0.25	0.52	0.44
sugars_100g	0.21	-0.58	-0.07	0.13	-0.15	0.53	-0.54	-0.03
fiber_100g	0.12	-0.06	0.81	-0.22	0.50	0.15	-0.04	-0.00
proteins_100g	0.12	0.38	0.43	-0.06	-0.77	0.19	-0.05	0.14
salt_100g	-0.04	0.15	0.20	0.96	0.12	0.01	-0.00	-0.00

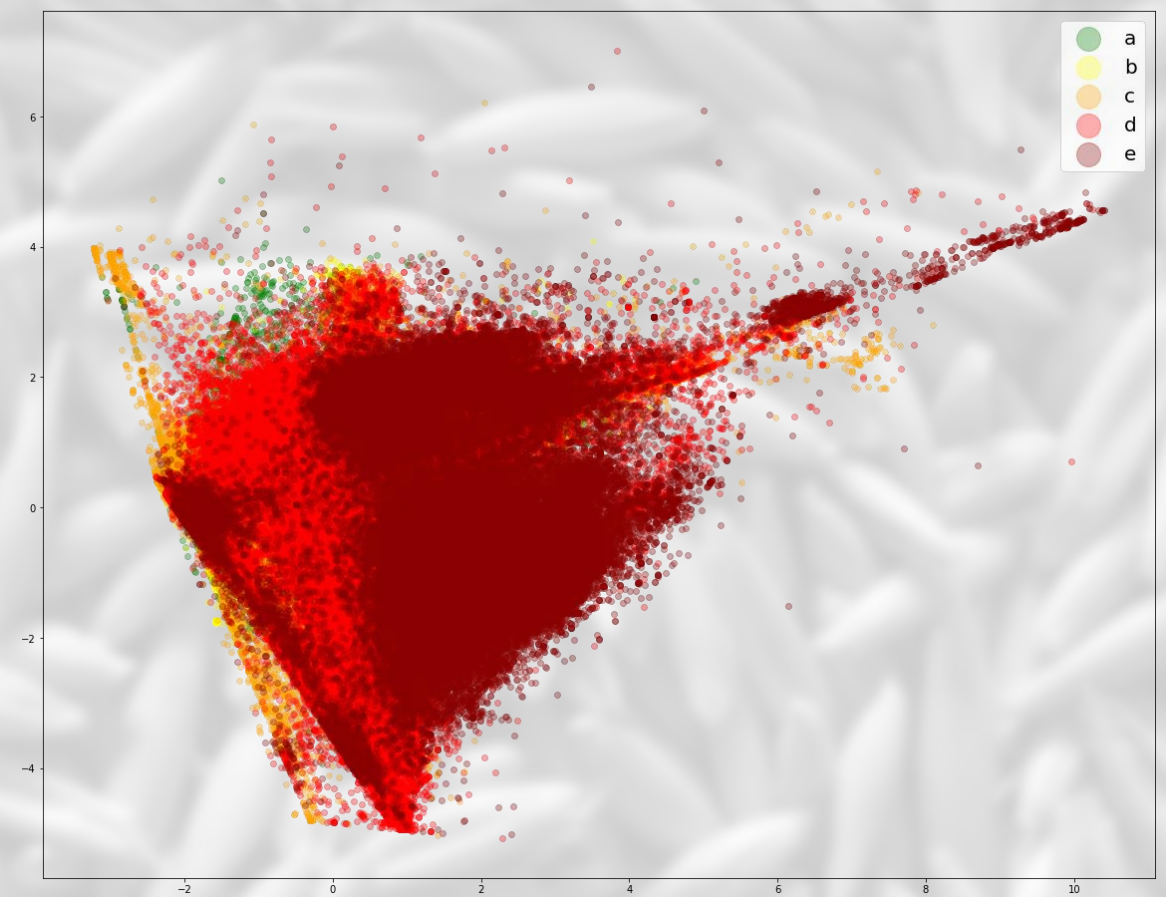
Pas de coude franc.
Plan factoriel 1 et 2 : 55,9%

F1 : richesse nutritionnelle (gras et energie)
F2 : sucre
F3 : fibre
F4 : sel

Cercle des corrélations (F1 et F2)

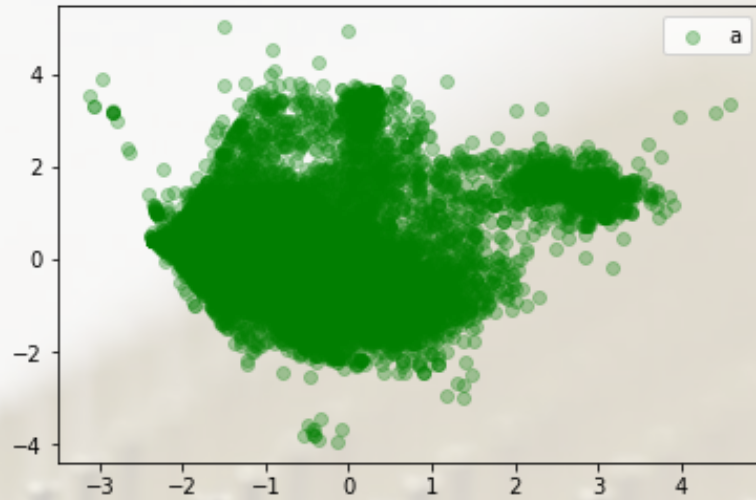


Représentation graphique de l'ACP sur les plans factoriels 1 et 2

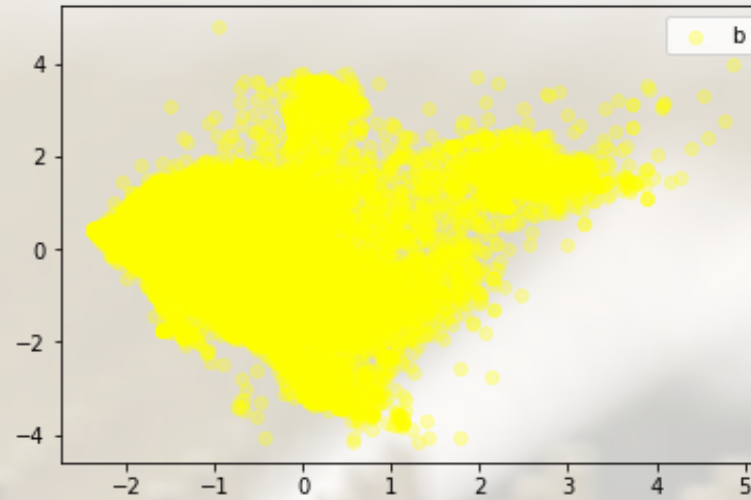


Représentations graphiques des plans factorielles 1 et 2

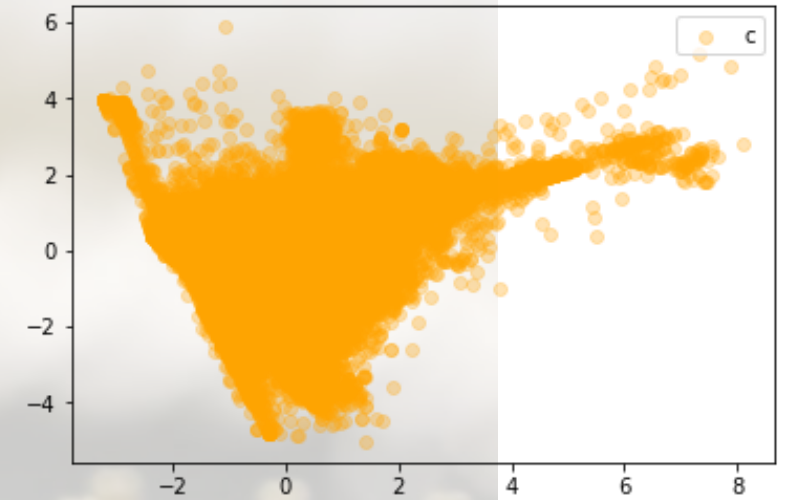
Représentation des produits de catégorie a sur le 1er plan factoriel



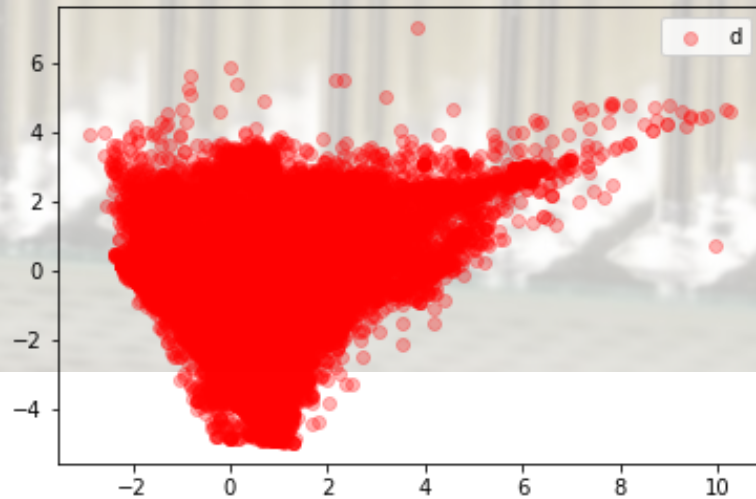
Représentation des produits de catégorie b sur le 1er plan factoriel



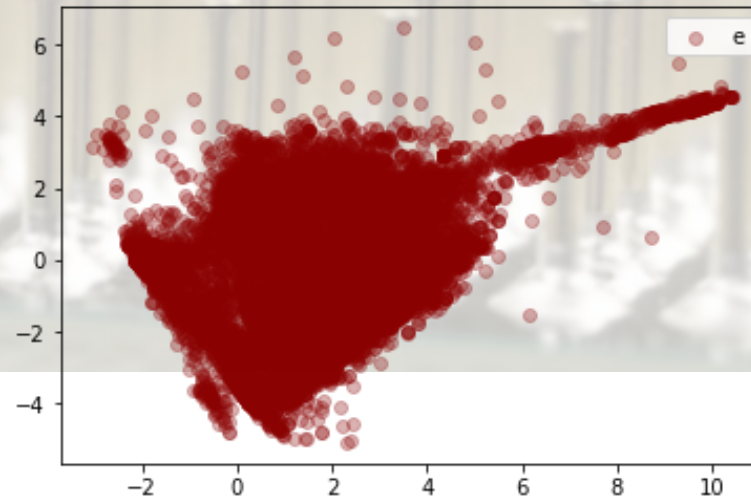
Représentation des produits de catégorie c sur le 1er plan factoriel



Représentation des produits de catégorie d sur le 1er plan factoriel



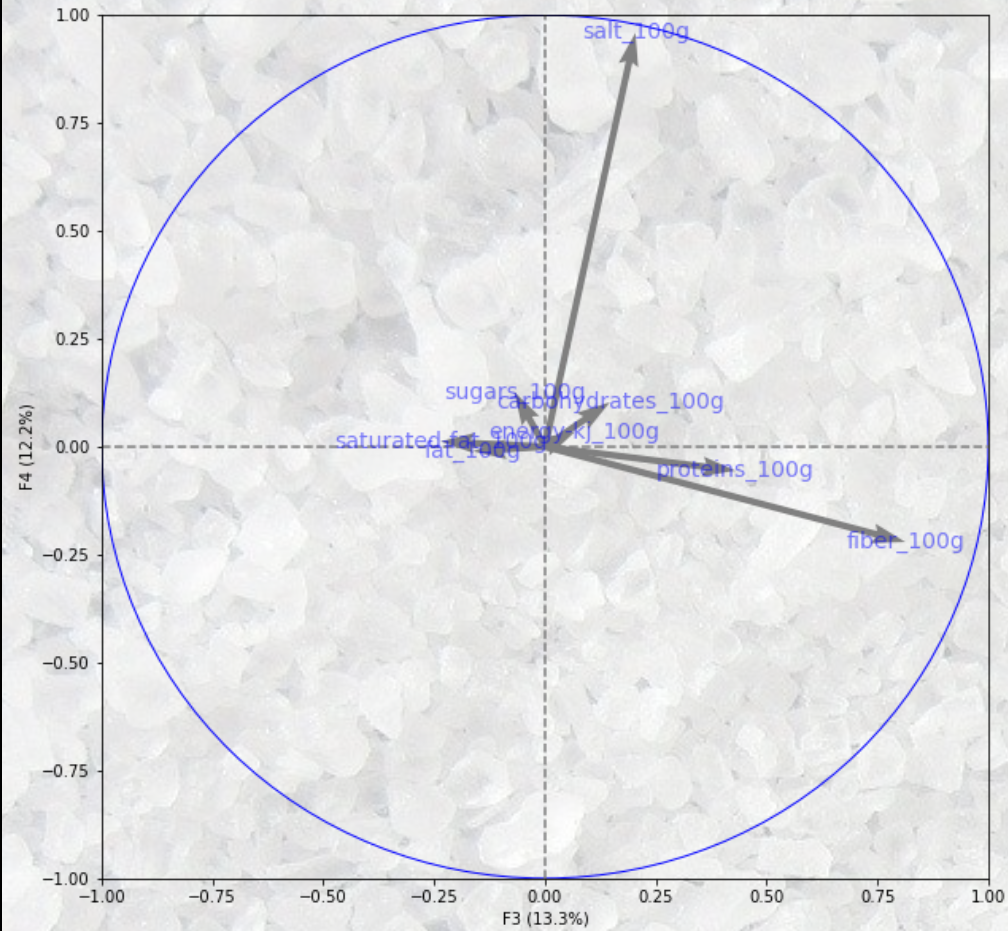
Représentation des produits de catégorie e sur le 1er plan factoriel



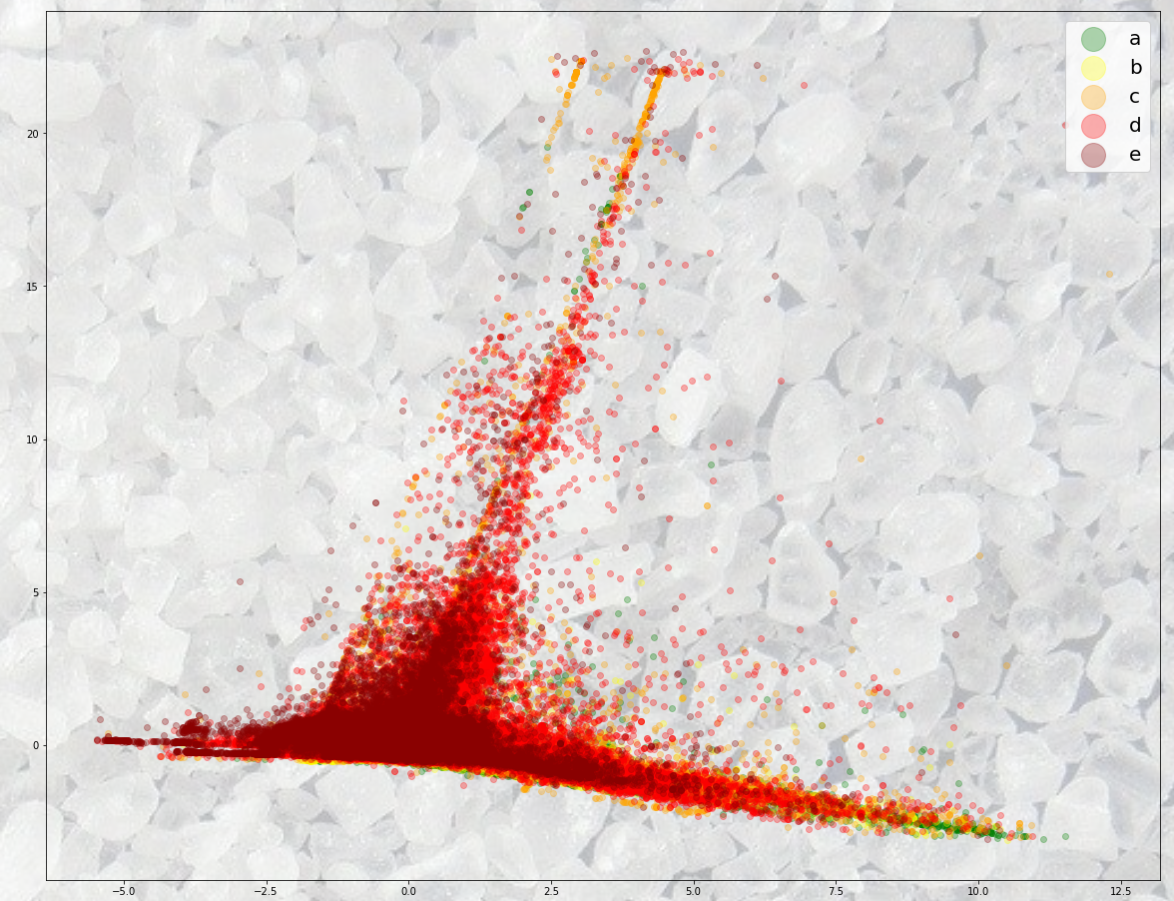
Les produits A sont +
recentrés mais
pas de distinction **nette**

La représentation factorielle
1 et 2 apparaît
insuffisante pour
expliquer le nutrigrade.

Cercle des corrélations (F3 et F4)

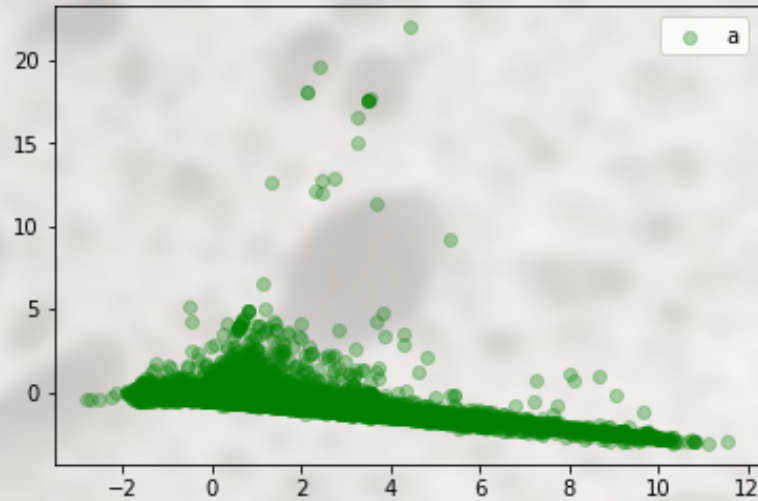


Représentation graphique de l'ACP sur les plans factoriels 3 et 4

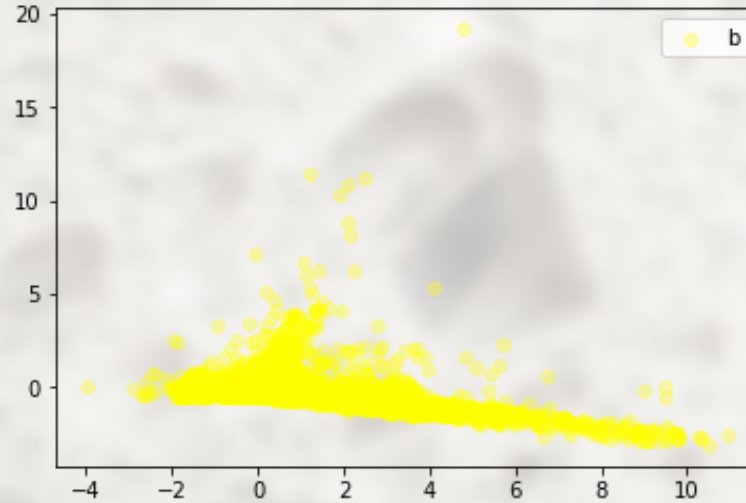


Représentations graphiques des plans factorielles 1 et 2

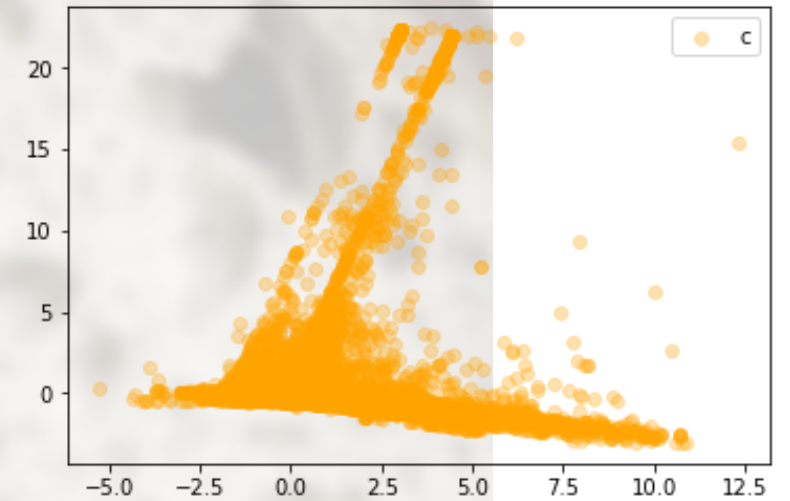
Représentation des produits de catégorie a sur le 2ème plan factoriel



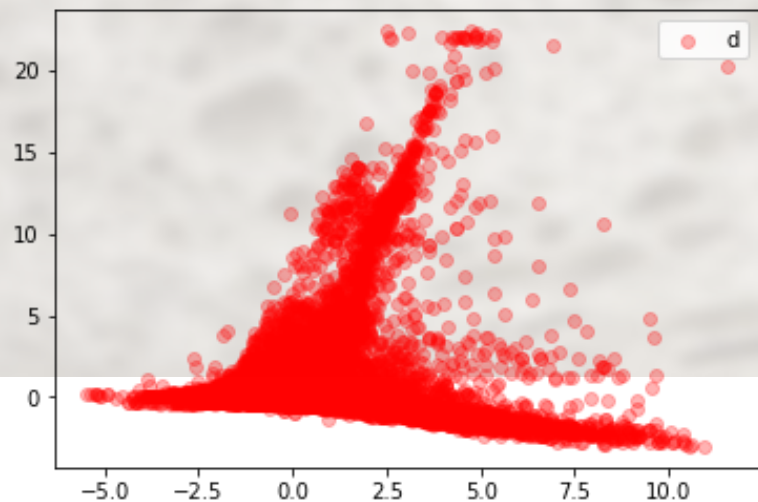
Représentation des produits de catégorie b sur le 2ème plan factoriel



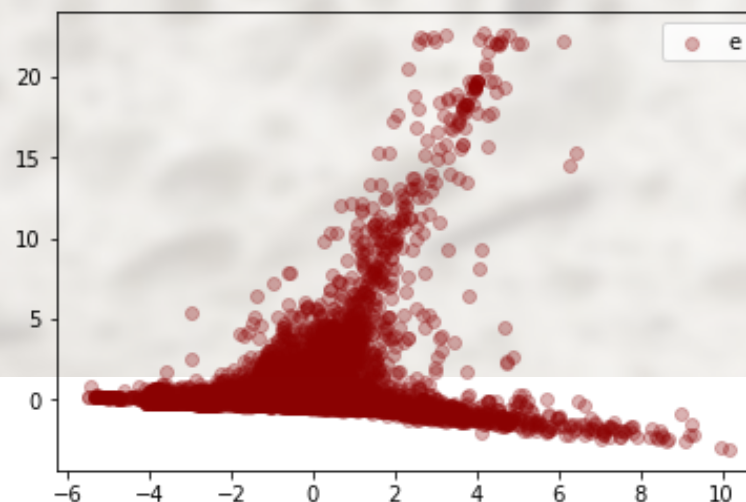
Représentation des produits de catégorie c sur le 2ème plan factoriel



Représentation des produits de catégorie d sur le 2ème plan factoriel



Représentation des produits de catégorie e sur le 2ème plan factoriel



Les produits **A et B** n'ont que rarement beaucoup de **sel**.



Là encore les catégories se **superposent**.

L'ACP confirme ce que les analyses bivariées avaient montré :
Les corrélations existent mais elles sont très floues.

ANOVA (analysis of variance)

 **62%** du nutriscore est expliqué par les variables nutritionnelles.

Ce score devrait être de 100% car le nutriscore est calculé **exclusivement** d'après les valeur nutritionnelles.

  La forte valeur du coefficient de Fischer et la faiblesse de la 'p-value' nous indiquent que le résultat de l'ANOVA n'est **pas du au hasard**.

OLS Regression Results

Dep. Variable:	nutriscore	R-squared:	0.620			
Model:	OLS	Adj. R-squared:	0.620			
Method:	Least Squares	F-statistic:	1.180e+05			
Date:	Thu, 19 Aug 2021	Prob (F-statistic):	0.00			
Time:	13:20:18	Log-Likelihood:	-1.7839e+06			
No. Observations:	579132	AIC:	3.568e+06			
Df Residuals:	579123	BIC:	3.568e+06			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.5736	0.017	148.704	0.000	2.540	2.608
energy_kj_100g	0.0033	2.56e-05	127.074	0.000	0.003	0.003
fat_100g	0.0250	0.001	22.604	0.000	0.023	0.027
saturated_fat_100g	0.4262	0.001	321.169	0.000	0.424	0.429
sugars_100g	0.1811	0.001	358.280	0.000	0.180	0.182
fiber_100g	-0.6398	0.005	-137.808	0.000	-0.649	-0.631
proteins_100g	0.0475	0.001	54.753	0.000	0.046	0.049
salt_100g	0.2792	0.002	171.105	0.000	0.276	0.282
carbohydrates_100g	-0.0327	0.001	-60.549	0.000	-0.034	-0.032
Omnibus:	29655.689	Durbin-Watson:	1.112			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	61209.919			
Skew:	-0.362	Prob(JB):	0.00			
Kurtosis:	4.419	Cond. No.	3.49e+03			

5

3 observation sur la pertinence et la faisabilité du projet



Faisabilité : **oui**
simple **techniquement**

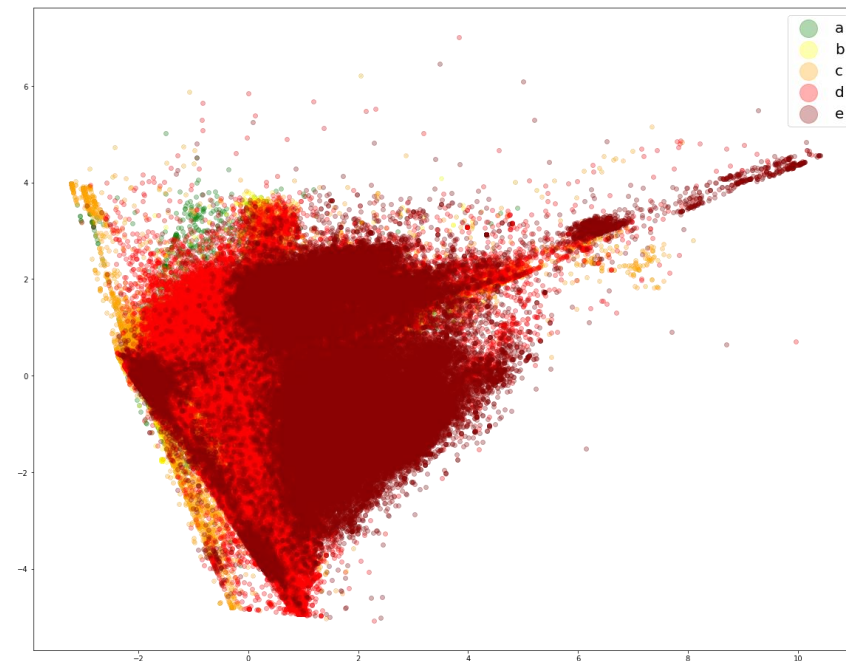
Pertinence : **faible**

les données même nettoyées sont de **mauvaise qualité**

ANOVA max score = **62%** alors que devrait être **100%**

Nutriscore ne s'applique pas aux produits non-transformés, pas aux produits pour bébé

Représentation graphique de l'ACP sur les plans factoriels 1 et 2



Dep. Variable:	nutriscore	R-squared:	0.620
Model:	OLS	Adj. R-squared:	0.620

	energy-kj_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	y_pred_R	y_pred_C	y_test
0	2198.0700	27.0	18.2	62.9	29.8	6.7	0.20	e	e	e
1	1256.0400	15.3	9.9	35.4	17.5	4.7	0.40	d	d	d
2	2122.7076	25.8	14.9	60.6	38.9	6.5	0.46	e	e	e
3	136.0000	0.5	0.1	5.4	5.4	10.0	0.50	a	a	e
4	598.0000	5.6	2.0	0.5	0.5	23.0	1.80	d	d	c



Conclusion :
**Problème de
qualité des
données**

Pertinence serait une base dont le contenu est alimenté par les industriels eux-mêmes.

À défaut, une application qui puisse lire les étiquettes des ingrédients et nutritionnelles des produits directement.

