Soutenance Projet 2

L'application de Nutrition pour le voyage

I Présentation de l'application

Scan le code barre d'un produit

soit le produit existe déjà dans les données et l'application retourne le Nutriscore

soit le produit n'existe pas et on peut remplir les nutriments principaux et l'application renvoit un nutriscore estimé

L'application propose des produits équivalents avec de meilleurs nutriscores

L'application propose des produits équivalents en fonction de certaines variables rentré par l'utilisateur, par exemple l'ecoscore si celui ci est renseigné

II Nettoyage

```
# Je défini qu'en dessous de 13 variables, je ne considère pas le produit dans l'exploitation

data["Nombres"] = data.apply(lambda x: x.count(), axis=1)
data = data[data["Nombres"]>=13]
```

Entrée [ ]:
```
# Je crée une nouvelle variable que j'ai appelé 100g. Elle additionne tous les nutriments. Cette variable va me permettre de
# supprimé certain outlier. Les nutriments étant rempli pour 100g, si la variable 100g est plus grande que 100g, cela veut dire
# que le produit a été mal renseigné

data["100g"] = data["carbohydrates_100g"] + data["salt_100g"] + data["fat_100g"] + data["proteins_100g"]+ data["fiber_100g"]
```

Entrée [ ]:
```
# Je remplace les valeurs manquantes de fiber par la moyenne

mean_fiber = data['fiber_100g'].mean()
data['fiber_100g'] = [mean(t, mean_fiber) for t in data['fiber_100g']]
```

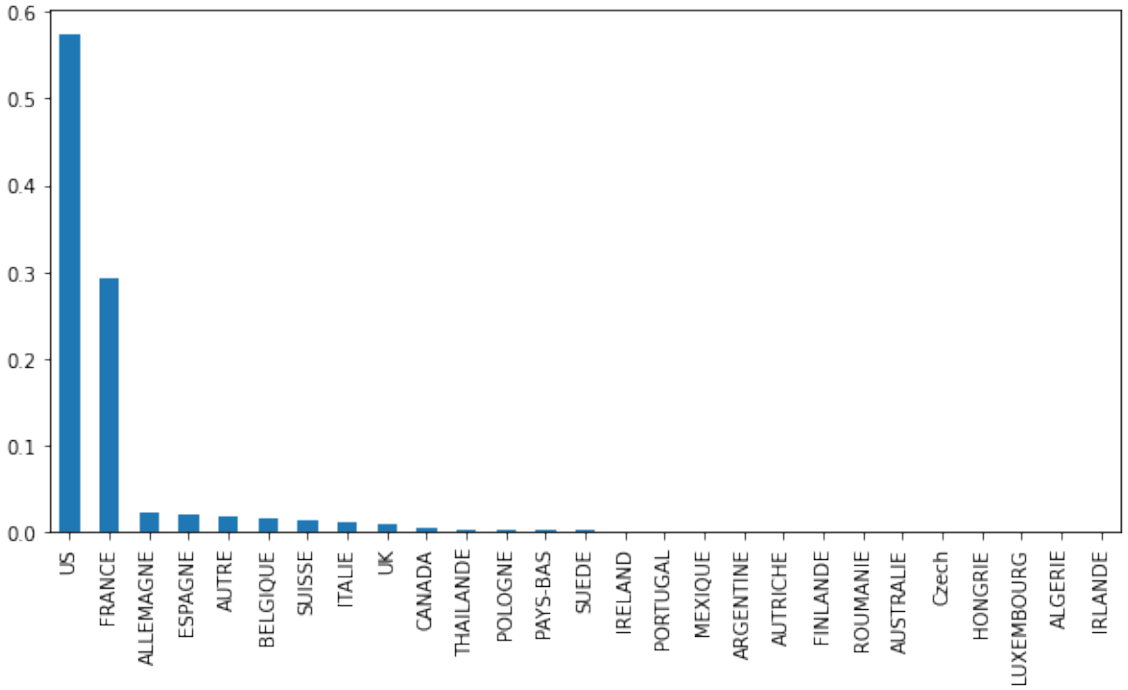
Entrée [ ]:
```
# Je supprime tous les outliers. Une valeur d'un nutriment ne peut pas dépasser 100g pour 100g de produit

data= data[data["sugars_100g"]<=100]
data= data[data["sodium_100g"]<=100]
data= data[data["proteins_100g"]<=100]
data= data[data["carbohydrates_100g"]<=100]
data= data[data["fat_100g"]<=100]
data= data[data["fiber_100g"]<=100]
data= data[data["salt_100g"]<=100]
data= data[data["saturated-fat_100g"]<=100]
```

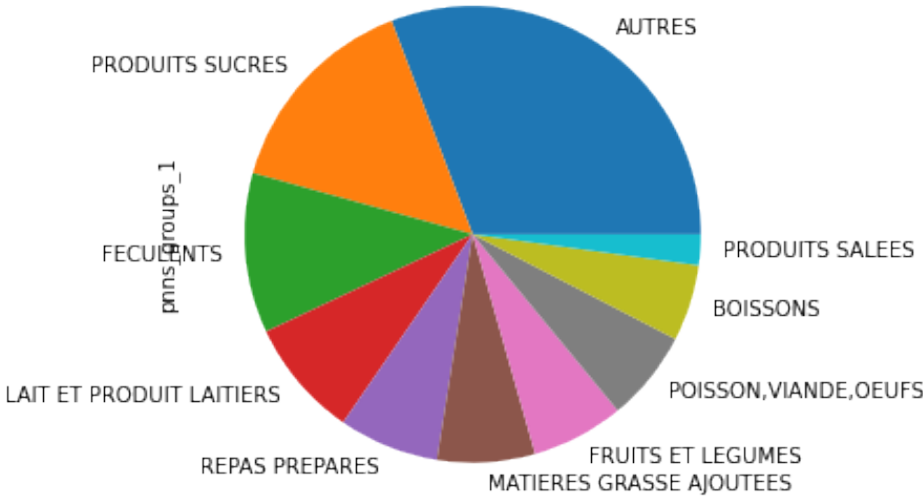
Entrée [ ]:
```
# J'applique ma nouvelle variable 100g

data= data[data["100g"]<=100]
```

Entrée [ ]:
```
for c in data.columns:
    if c in ['energy-kcal_100g', 'energy_100g']:
        subset = data[c] # Création du sous-échantillon
        #On calcule Q1

        q1=subset.quantile(q=0.25)

        #On calcule Q3

        q3=subset.quantile(q=0.75)

        #On calcule l'écart interquartile (IQR)

        IQR=q3-q1

        #On calcule la borne inférieure à l'aide du Q1 et de l'écart interquartile

        borne_inf = q1-1.5*IQR

        #On calcule la borne supérieure à l'aide du Q3 et de l'écart interquartile

        borne_sup = q3 +1.5*IQR

        #On garde les valeurs à l'intérieur de la borne inférieure et supérieure

        data= data[data[c]<borne_sup]
        data =data[data[c]>borne_inf]
```

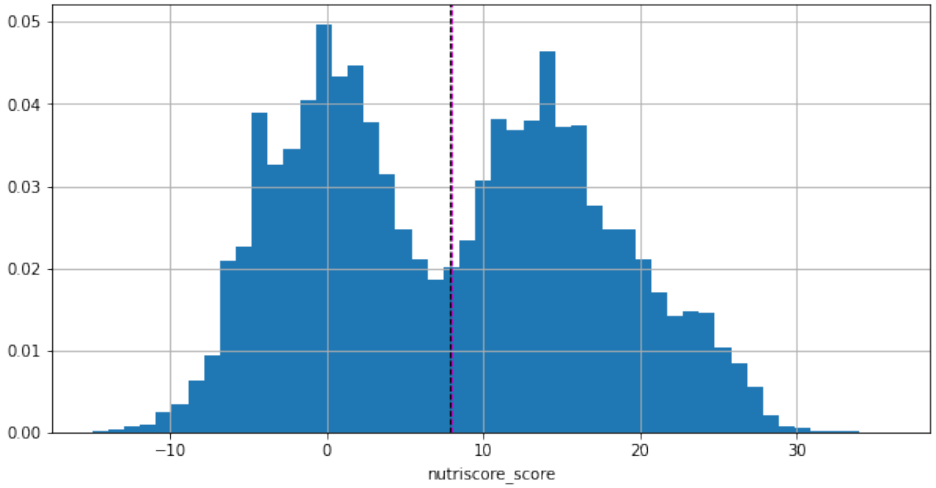
II Nettoyage

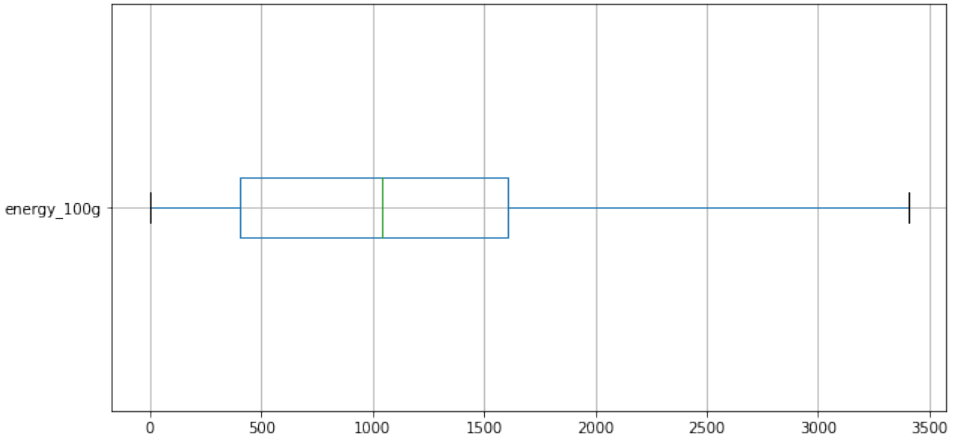
Après le nettoyage, cela me permet d'avoir des données sur les variables qualitatives, comme le pays de provenance du produit et les différents types de produits classées selon le pnns

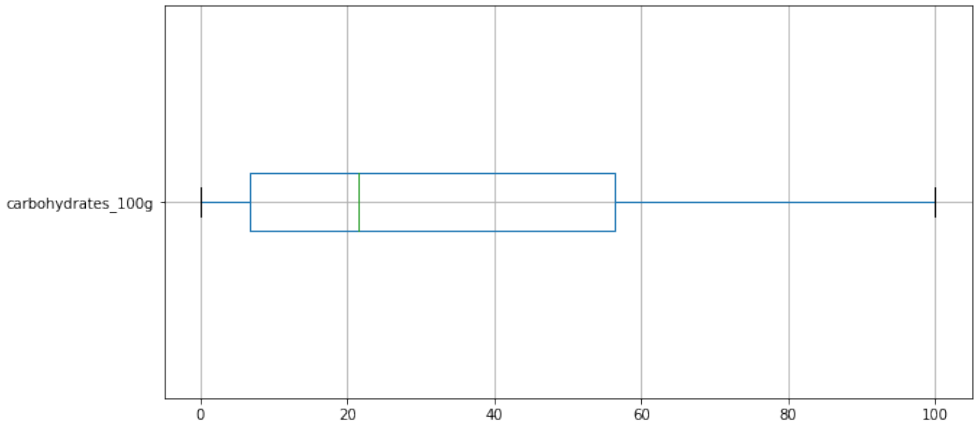


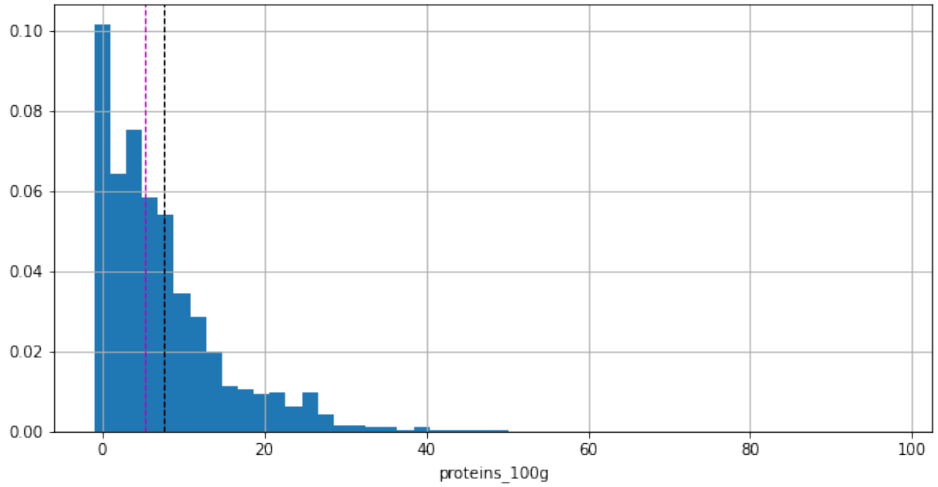


III Exploration (univarié)



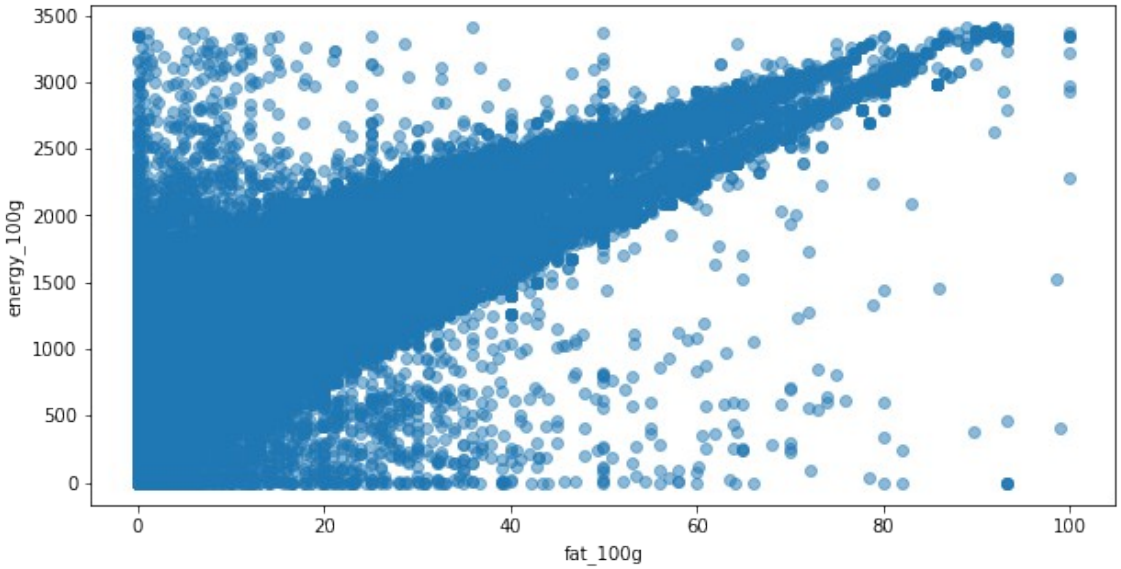




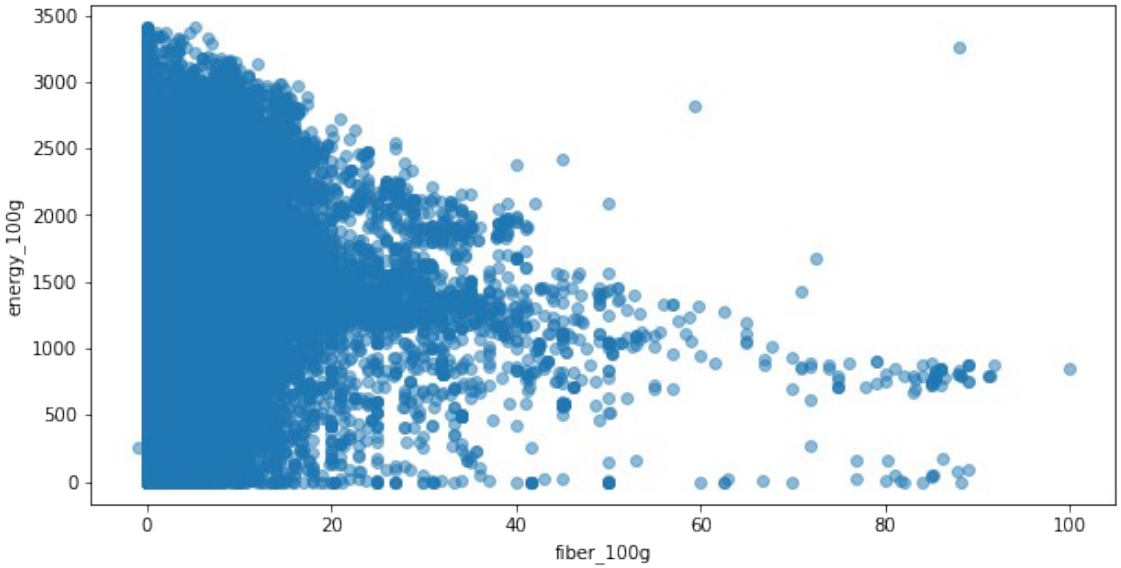


III Exploration (multivarié)

-L'énergie est corrélée aux matières grasses, plus on augmente les matières grasses plus l'on a d'énergies en kJ

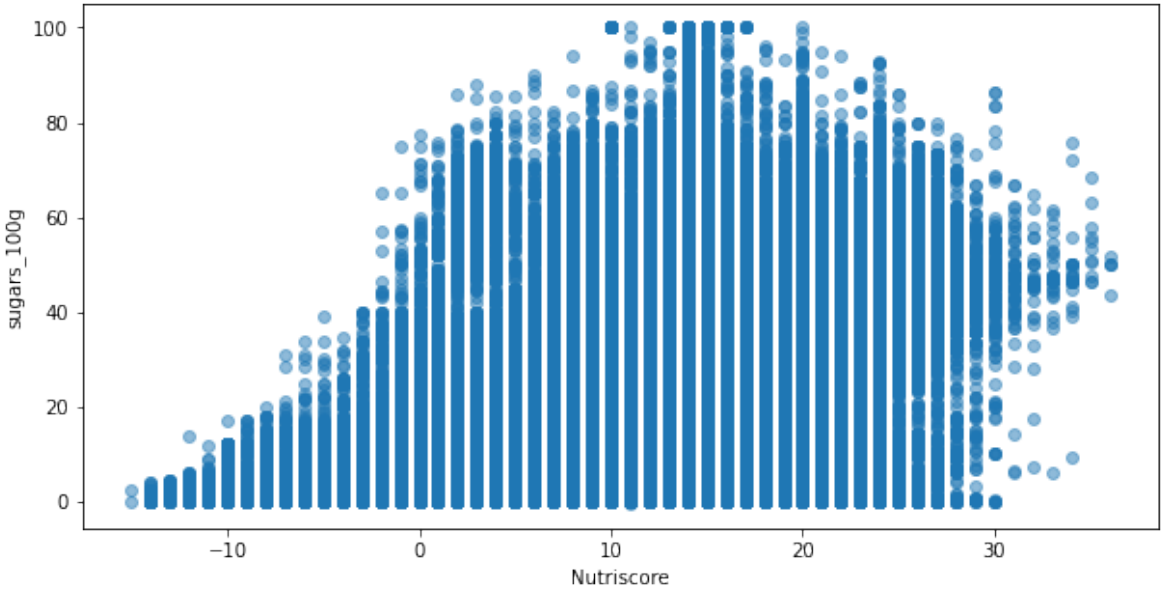


-L'énergie est corrélée aux fibres, plus on augmente les fibres moins l'énergie est grande en kJ

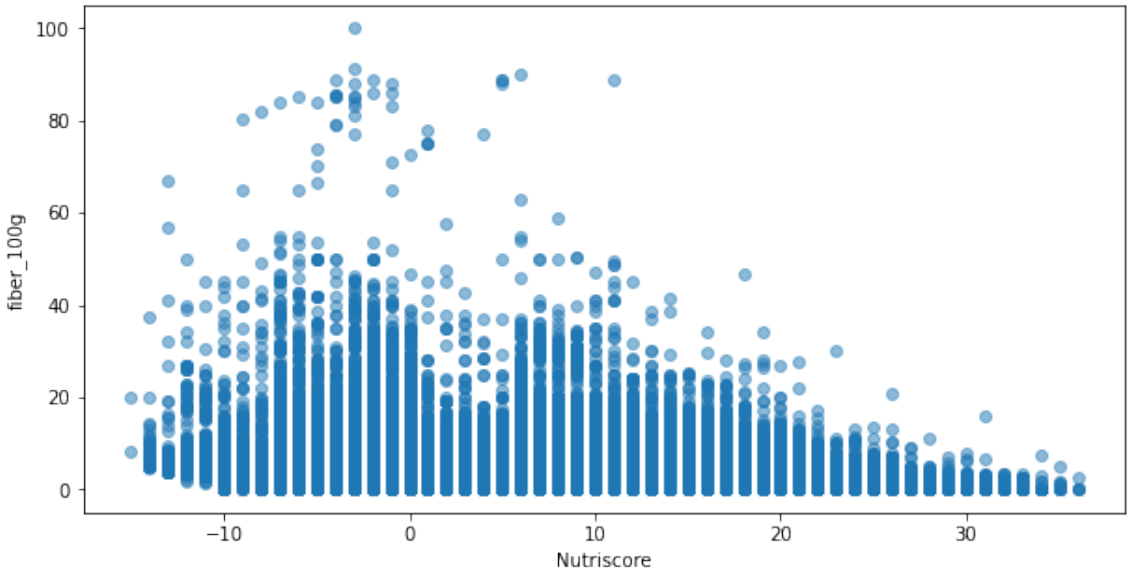


III Exploration (multivarié)

-Le Nutriscore est corrélé aux sucres, plus on augmente les sucres plus le Nutriscore est élevé

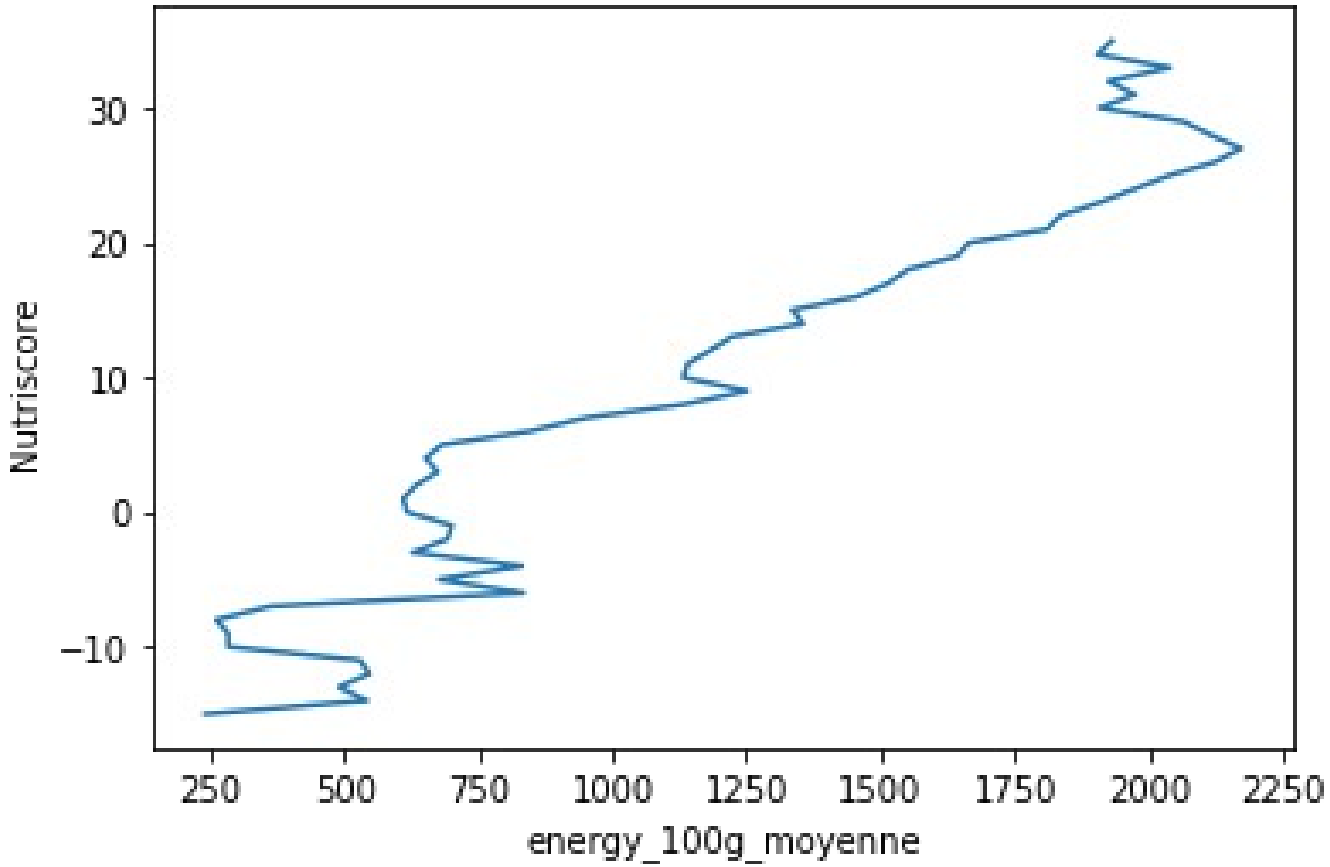


-Le Nutriscore est corrélé aux fibres, plus on augmente les fibres moins le Nutriscore est élevé



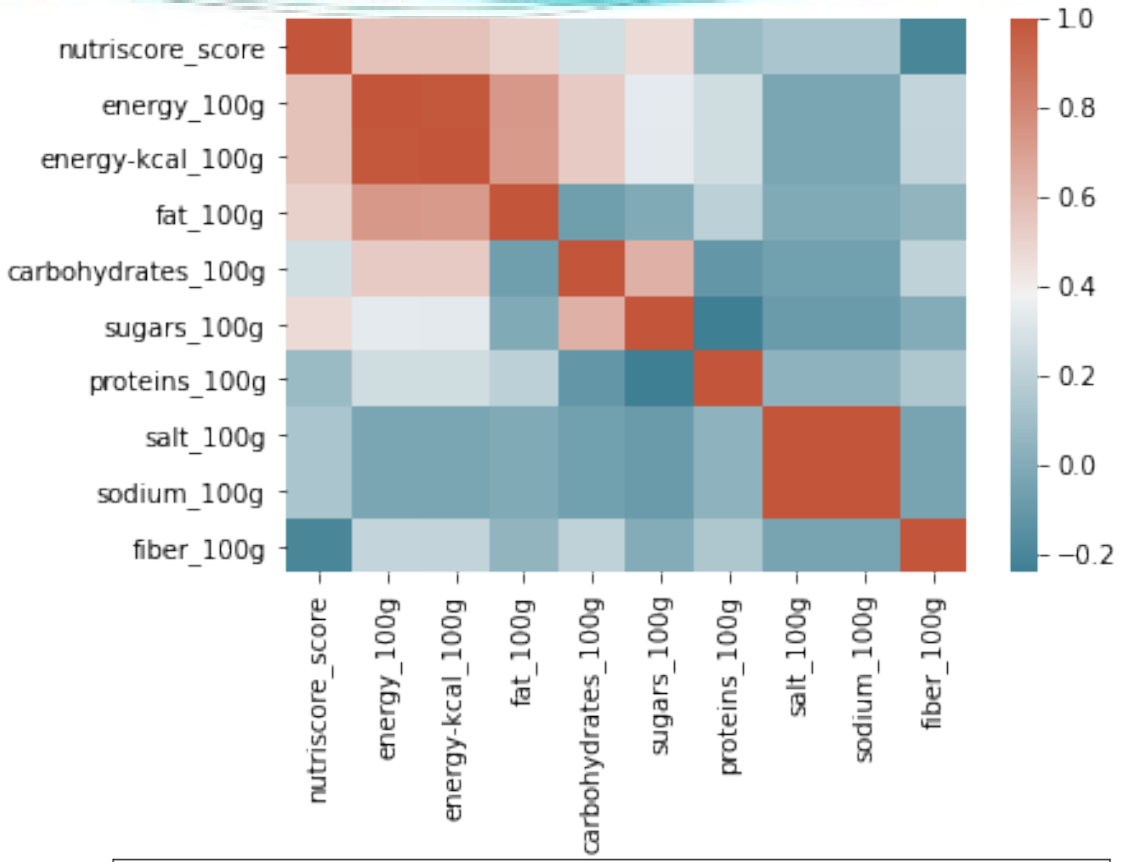
III Exploration (multivarié)

Grâce à nos études précédentes, on peut supposer qu'il existe une corrélation entre l'énergie et le nutriscore. Pour mieux représenter celle-ci, j'ai décidé de prendre le nutriscore en fonction de l'énergie moyenne



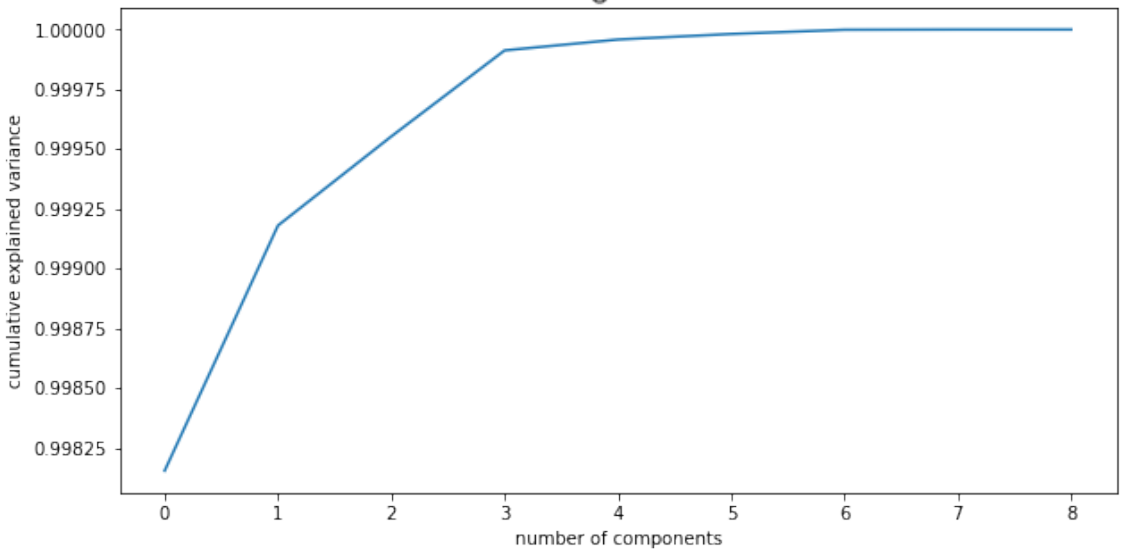
III Exploration (réduction de données)

J'ai donc effectué une heatmap des corrélations pour mieux rendre compte de ce que l'on a pu observer grâce aux analyses multivariées précédente.



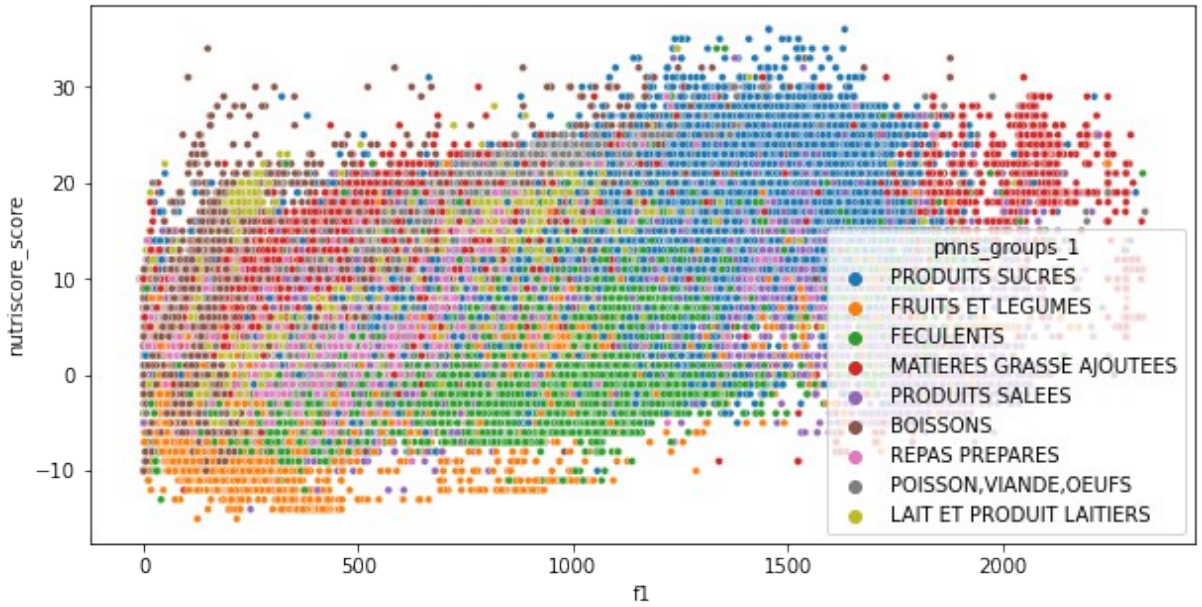
Ces corrélations nous a amené à nous poser la question d'une réduction des variables nutriments.

Graphique pour déterminer le nombre de composantes principales.



IV Application

On peut repérer grâce à notre nouvelle variable, on peut observer des zones différentes de Nutriscore pour les classifications des produits.



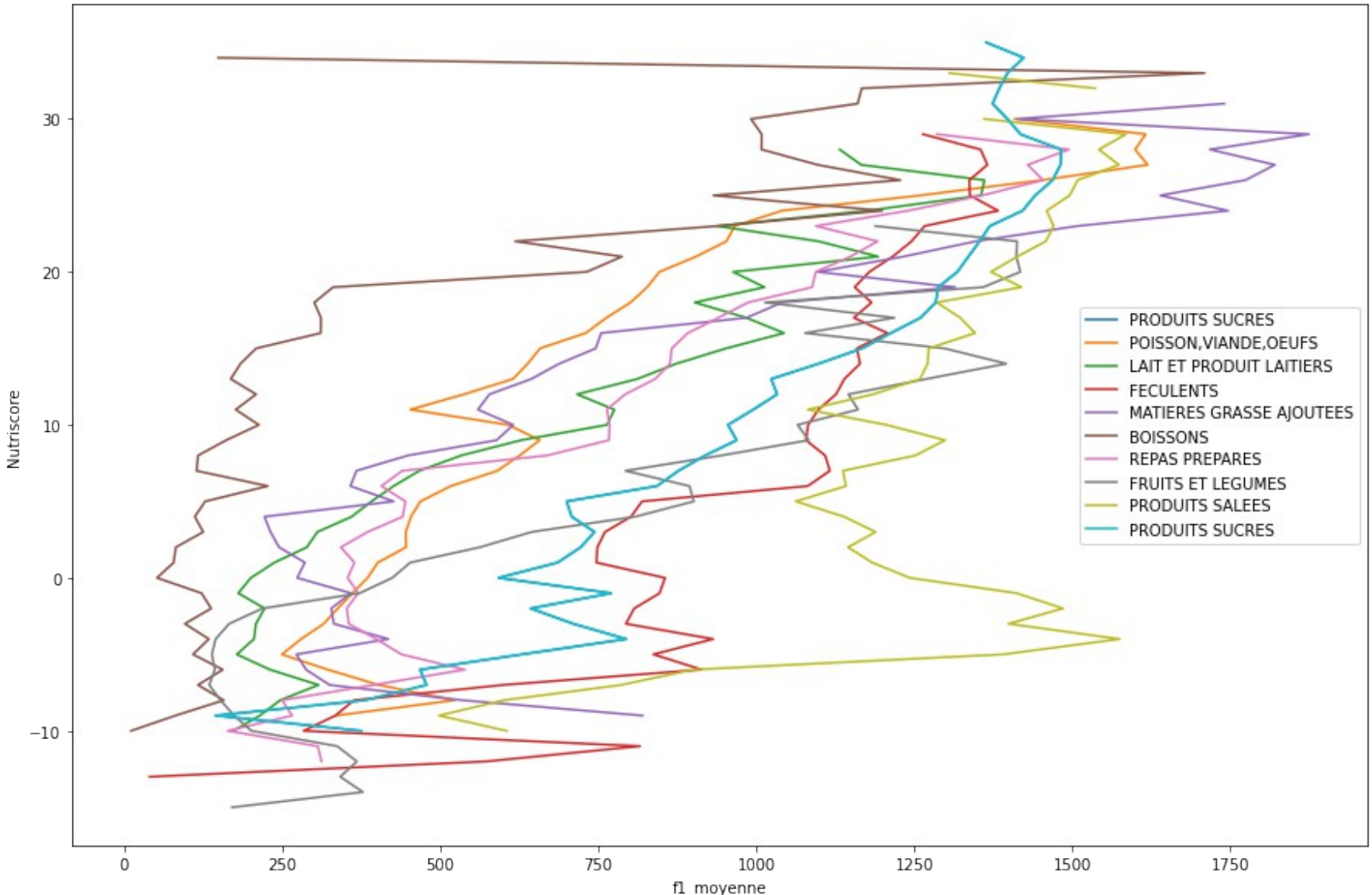
Il existe une deuxième variables qualitatives où la classification des produits est plus précise. Ceci peut nous permettre de mieux les différencier. C'est un bon éléments pour notre futur applications.

	pnns_groups_1	pnns_groups_2	nutriscore_score
0	AUTRE	unknown	-13.0
1	BOISSONS	Fruit juices	-10.0
2	FECULENTS	Potatoes	-13.0
3	FRUITS ET LEGUMES	Vegetables	-15.0
4	LAIT ET PRODUIT LAITIERS	Milk and yogurt	-10.0
5	MATIERES GRASSE AJOUTEES	Dressings and sauces	-12.0
6	POISSON,VIANDE,OEUFS	Fish and seafood	-9.0
7	PRODUITS SALEES	Appetizers	-14.0
8	PRODUITS SUCRES	Biscuits and cakes	-10.0
9	REPAS PREPARES	One-dish meals	-12.0

	pnns_groups_1	pnns_groups_2	nutriscore_score
0	AUTRE	unknown	36.0
1	BOISSONS	Sweetened beverages	34.0
2	FECULENTS	Bread	34.0
3	FRUITS ET LEGUMES	Soups	23.0
4	LAIT ET PRODUIT LAITIERS	Milk and yogurt	34.0
5	MATIERES GRASSE AJOUTEES	Dressings and sauces	31.0
6	POISSON,VIANDE,OEUFS	Fish and seafood	30.0
7	PRODUITS SALEES	Appetizers	33.0
8	PRODUITS SUCRES	Sweets	36.0
9	REPAS PREPARES	One-dish meals	29.0

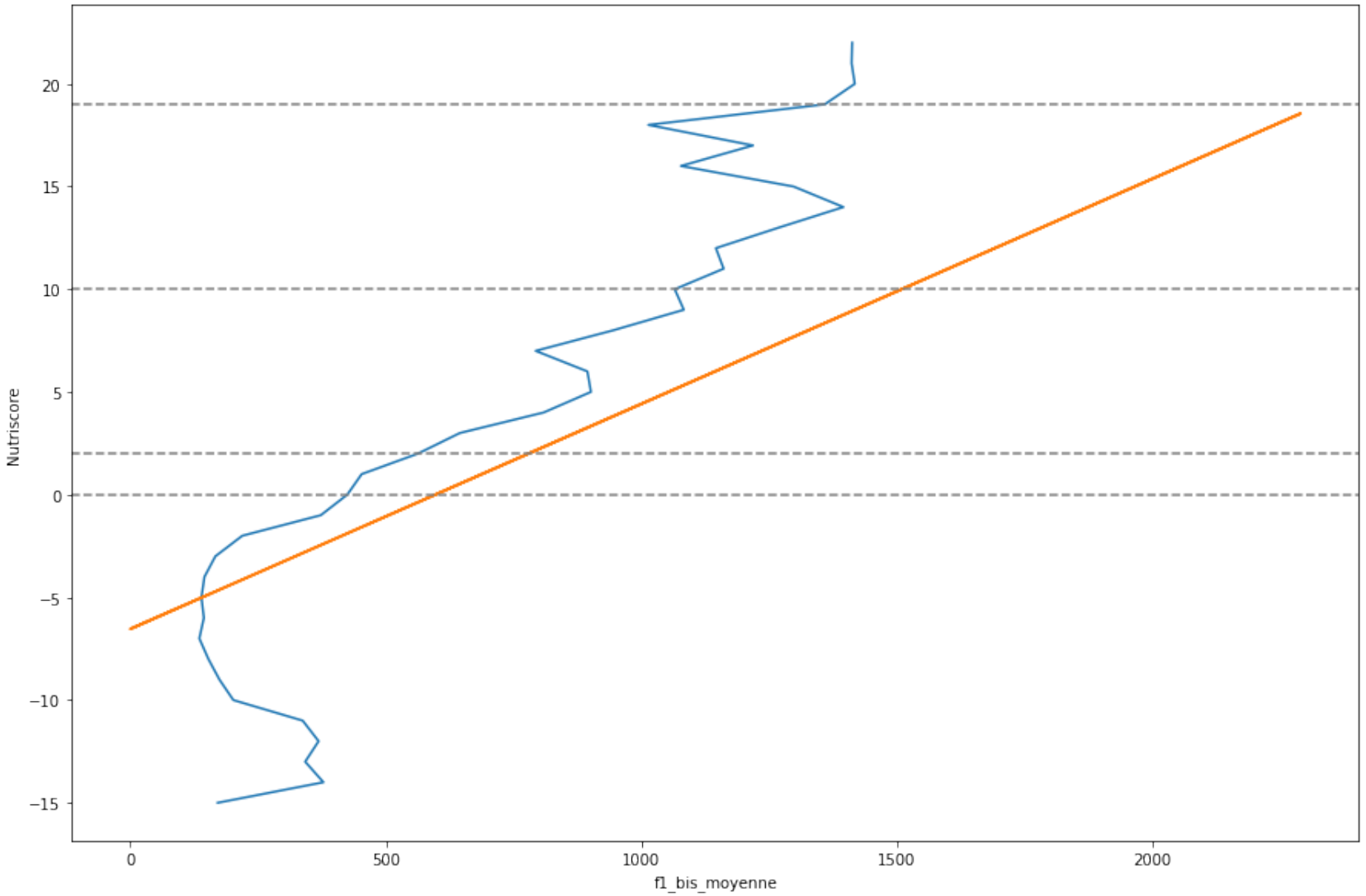
IV Application

On peut également tracer le nutriscore en fonction de la variable f1 moyenne. Cette fois-ci, j'ai tracé en fonction de la classe du produit. On peut remarquer que, pour certaines, le nutriscore a l'air d'avoir une dépendance linéaire en f1 comme attendu.



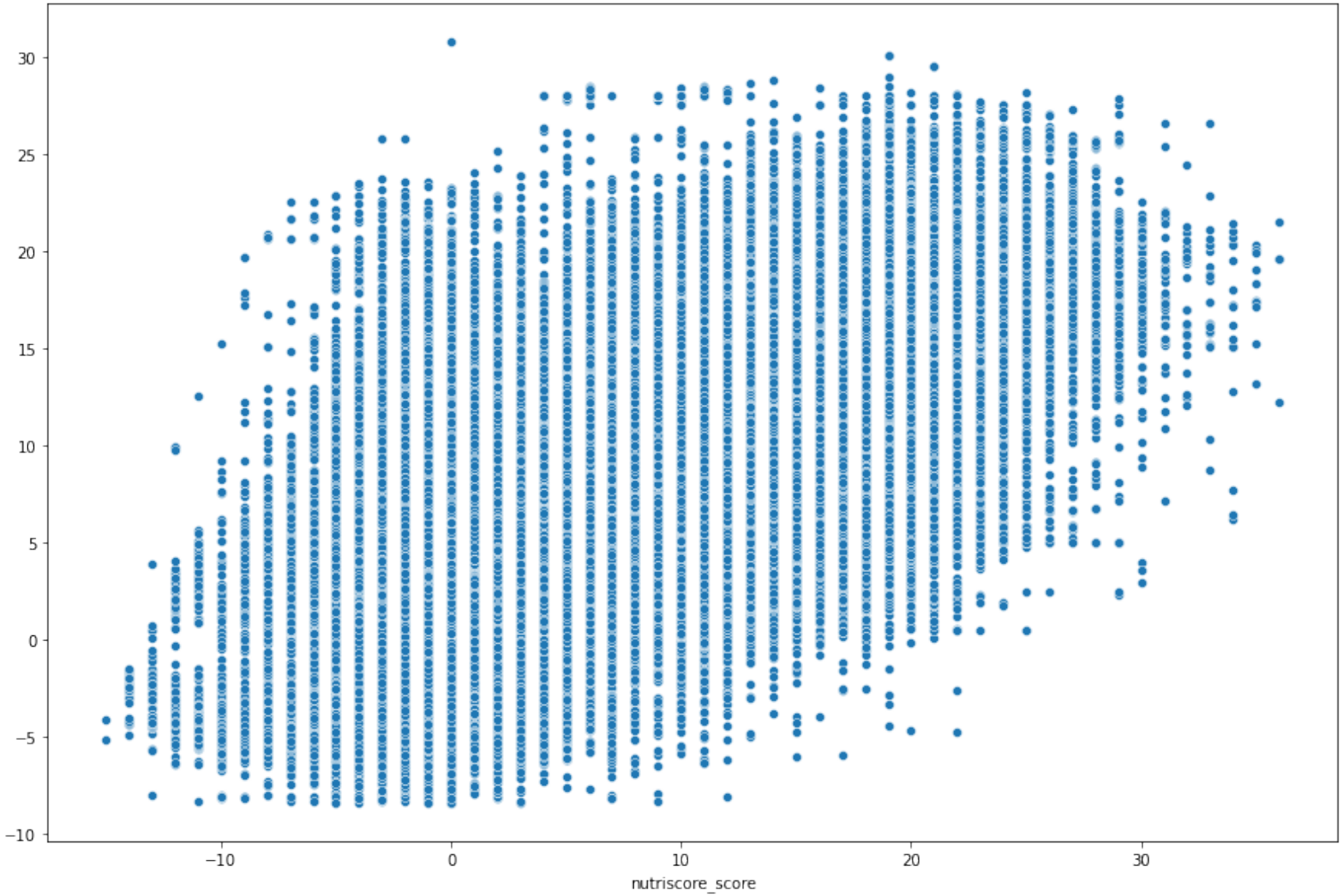
IV Application

Pour mieux rendre compte de la possibilité de notre application, j'ai décidé de tester une régression linaire sur une des classes. Les lignes pointillées correspondent aux limites entre les différentes lettres du nutriscore.



IV Application

Malheureusement, il semble que la régression linéaire ne soit pas suffisante pour estimer le nutriscore. Ce que l'on peut voir, c'est une grande différences, entre le nutriscore réel et le nutriscore estimé. C'est confirmé par le R^2 de 0,59.



Merci de votre attention