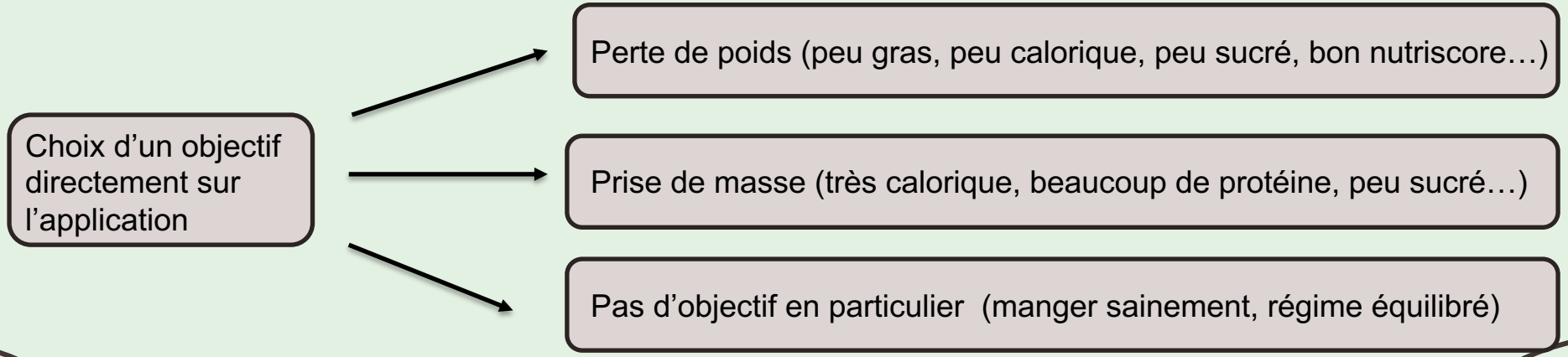
Santé publique France

Concevez une application au service de la santé publique

Sommaire

Introduction : Présentations de l'Application

- Application destinée aux sportifs qui veulent manger sainement.



Introduction : Présentations de l'Application

FoodY

Scan du produit

Donne un score sur 100 du produit

Non satisfait

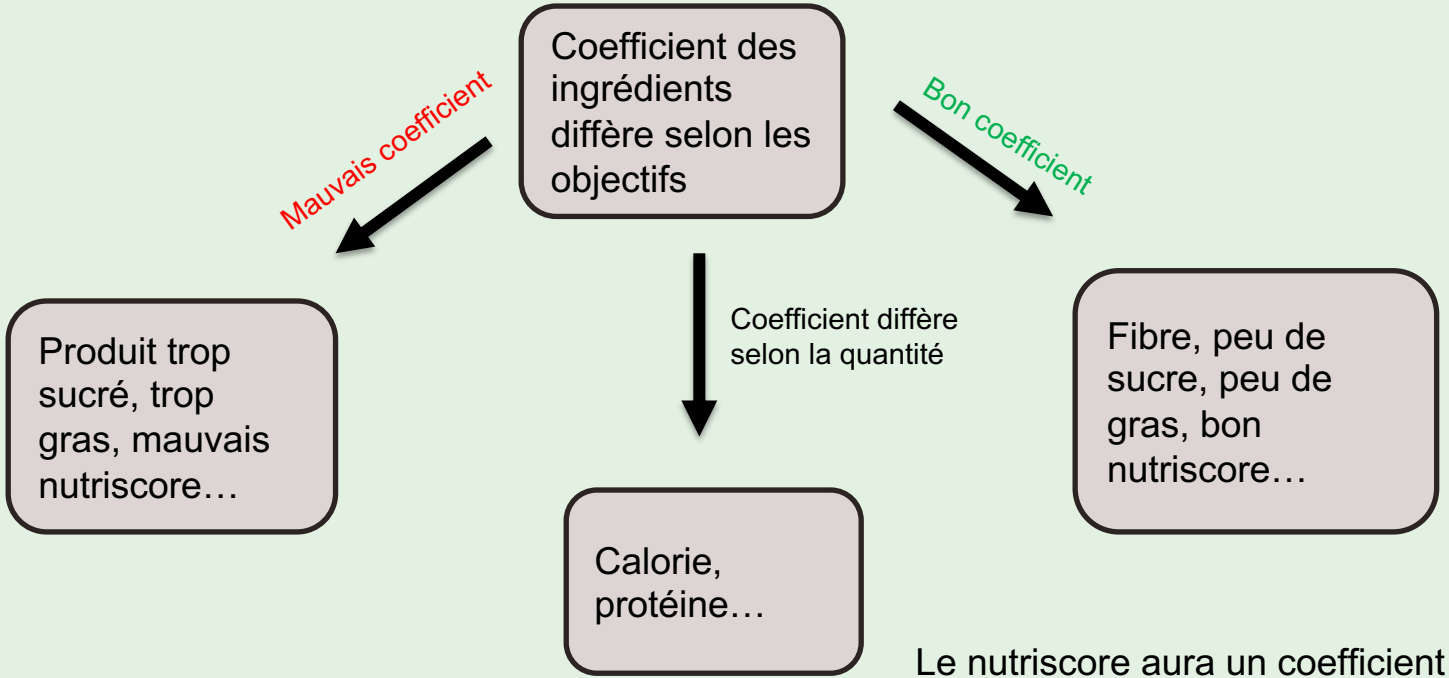
Propose des produits similaire mieux noté et avec un meilleur nutriscore

Satisfait

Satisfait

Fin du processus

Comment fonctionne la notation ?



Le nutriscore aura un coefficient élever

Analyse exploratoire des données

Présentation du document :

- 320772 lignes pour 162 colonnes
- 5 groupes de variable différente parmi les colonnes

→

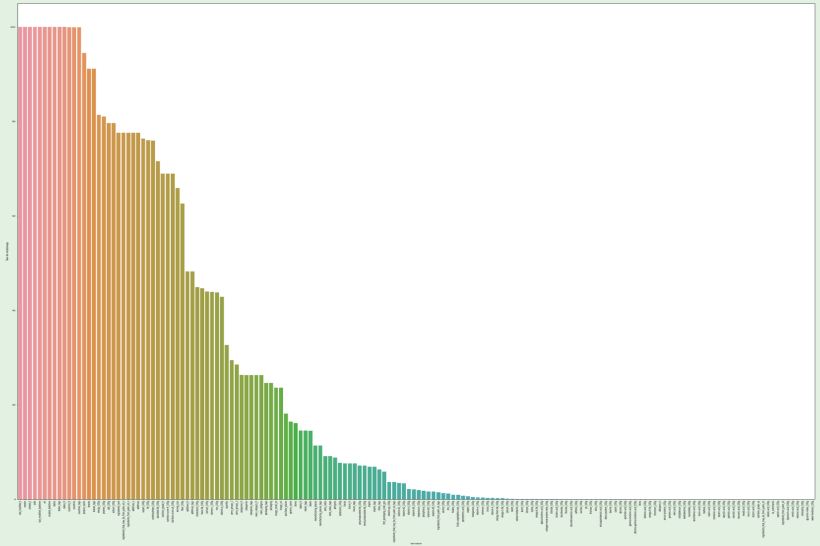
- Informations générales
- Tags
- Ingrédients
- Informations nutritionnelles
- Données diverses

Nettoyage

Un total de 39608627 données manquante

Beaucoup de colonnes entièrement vides ou très peu de données

Objectif : sélectionner les colonnes utiles à notre analyse



Taux de remplissage des colonnes

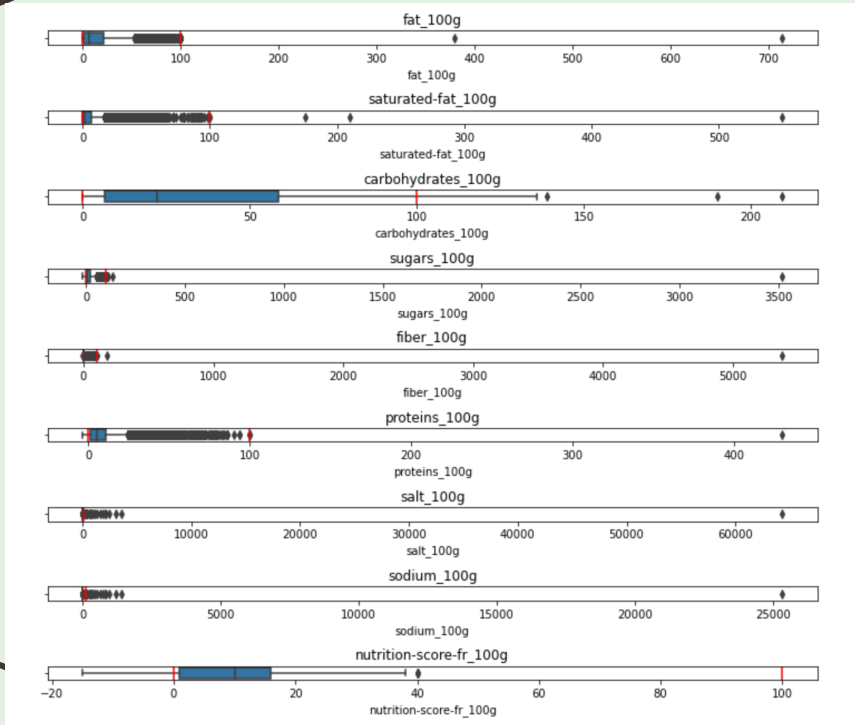
Sélections des colonnes utile a l’analyse

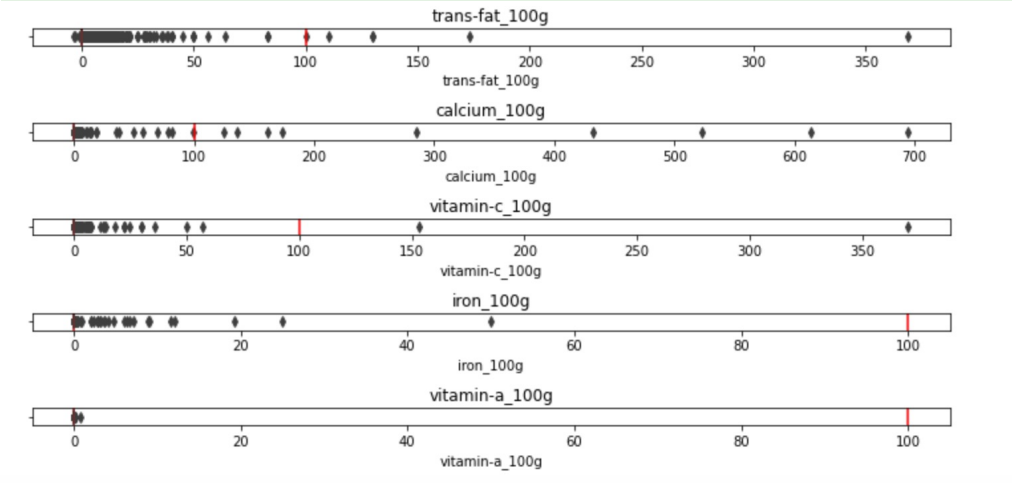
1 - Sélections des colonnes ayant un taux de remplissage supérieur a 60 %

2 - Sélections des colonnes utile a notre application

3 – Supprimer les lignes ayant un taux de remplissage inférieur a 60%

Outliers


fat_100g
0 100 200 300 400 500 600 700
fat_100g
saturated-fat_100g
0 100 200 300 400 500
saturated-fat_100g
carbohydrates_100g
0 50 100 150 200
carbohydrates_100g
sugars_100g
0 500 1000 1500 2000 2500 3000 3500
sugars_100g
fiber_100g
0 1000 2000 3000 4000 5000
fiber_100g
proteins_100g
0 100 200 300 400
proteins_100g
salt_100g
0 10000 20000 30000 40000 50000 60000
salt_100g
sodium_100g
0 5000 10000 15000 20000 25000
sodium_100g
nutrition-score-fr_100g
−20 0 20 40 60 80 100
nutrition-score-fr_100g


trans-fat_100g
0 50 100 150 200 250 300 350
trans-fat_100g
calcium_100g
0 100 200 300 400 500 600 700
calcium_100g
vitamin-c_100g
0 50 100 150 200 250 300 350
vitamin-c_100g
iron_100g
0 20 40 60 80 100
iron_100g
vitamin-a_100g
0 20 40 60 80 100
vitamin-a_100g

	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g	cholesterol_100g	trans-fat_100g	calcium_100g	vitamin-c_100g	iron_100g	vitamin-a_100g
count	209810.0	202619.0	209811.0	204995.0	176979.0	210156.0	210548.0	210543.0	196489.0	143481.0	142328.0	140094.0	139941.0	140026.0	137124.0
mean	13.5	5.0	32.6	15.3	2.9	7.6	1.8	0.7	9.2	0.0	0.1	0.1	0.0	0.0	0.0
std	17.5	7.8	28.5	21.5	13.5	8.2	141.1	55.6	9.1	0.4	1.5	3.3	1.1	0.2	0.0
min	0.0	0.0	0.0	-17.9	-6.7	-3.6	0.0	0.0	-15.0	0.0	-3.6	0.0	-0.0	-0.0	-0.0
25%	0.5	0.0	6.7	1.4	0.0	1.3	0.1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
50%	6.7	1.8	22.2	5.4	1.6	5.3	0.6	0.3	10.0	0.0	0.0	0.0	0.0	0.0	0.0
75%	21.4	7.1	58.6	23.3	3.6	10.7	1.4	0.6	16.0	0.0	0.0	0.1	0.0	0.0	0.0
max	714.3	550.0	209.4	3520.0	5380.0	430.0	64312.8	25320.0	40.0	95.2	369.0	694.7	370.4	50.0	0.8

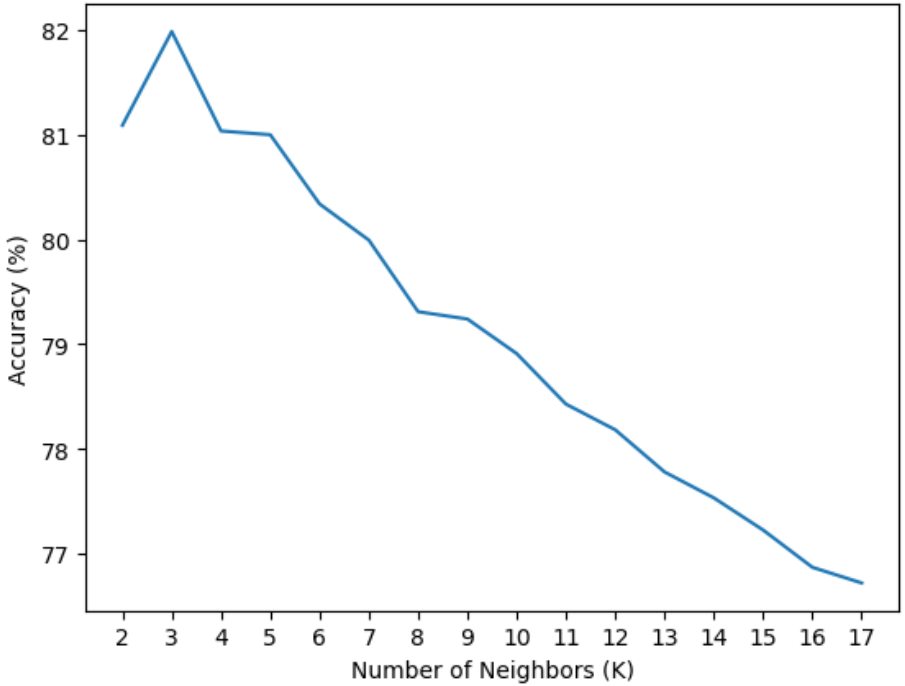
	energy_100g
count	210461.0
mean	1188.0
std	7153.0
min	0.0
25%	434.0
50%	1172.0
75%	1700.0
max	3251373.0

- 32247 valeurs inférieures a 0, remplacer par 0.
- 121 valeurs supérieures a 100, supprimer.
- Pour Energy, 3 valeurs supérieures a 3765.5, supprimer.

Imputation des valeurs manquantes

L'imputation de KNN peut être efficace car elle utilise les données des observations les plus proches pour estimer les valeurs manquantes.

En utilisant l'algorithme KNeighbourClassifier, on obtient le meilleur nombre de voisins qui est de K=3




```
k = 3  accuracy = 81.98400412796698 %
```

Conclusion nettoyage

-193799 lignes et 24 colonnes

Entrée [63]:
```
dataKNN2
```

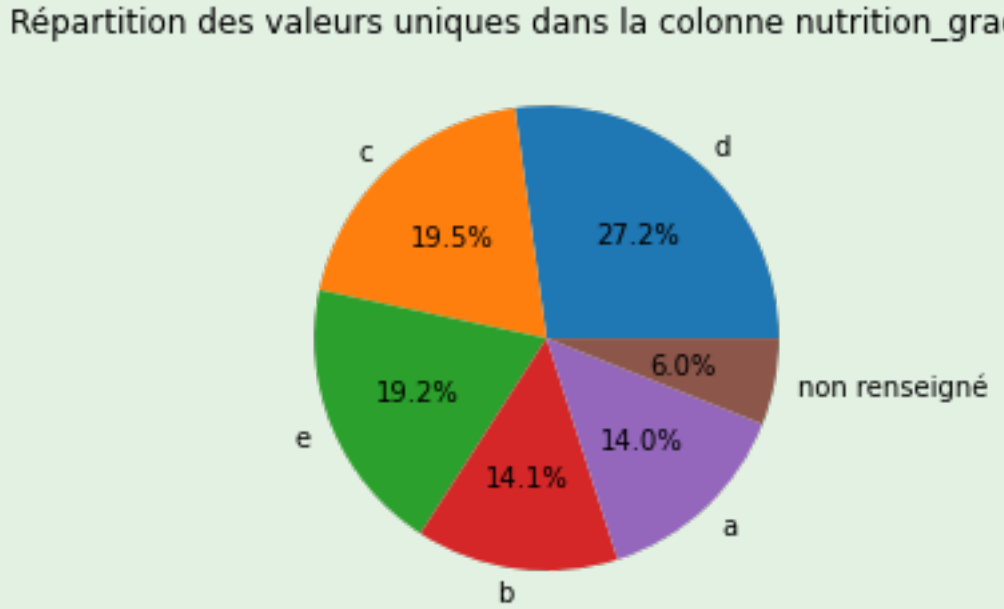
Out[63]:

	countries_fr	product_name	energy_100g	proteins_100g	salt_100g	sodium_100g	ingredients_that_may_be_from_palm_oil_n	ingredients_from_palm_oil_n
2	États-Unis	Peanuts	1941.0	17.9	0.6	0.2	0.0	0.0
3	États-Unis	Organic Salted Nut Mix	2540.0	17.9	1.2	0.5	0.0	0.0
5	États-Unis	Breadshop Honey Gone Nuts Granola	1933.0	13.5	0.9	0.3	0.0	0.0
7	États-Unis	Organic Muesli	1833.0	14.1	0.1	0.1	0.0	0.0
8	États-Unis	Organic Dark Chocolate Minis	2406.0	5.0	0.1	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
320693	Royaume-Uni	Santa Cruz Chilli & Lime Dressing	660.0	0.3	0.5	0.2	0.0	0.0
320702	France	Fisherman's Friend Miel-Citron	1031.0	0.0	0.0	0.0	1.0	0.0
320738	États-Unis	Organic Z Bar	1393.0	5.6	1.0	0.4	0.0	0.0
320742	États-Unis	Natural Cassava	1477.0	1.2	0.0	0.0	0.0	0.0
320763	France	Thé vert Earl grey	21.0	0.5	0.0	0.0	0.0	0.0

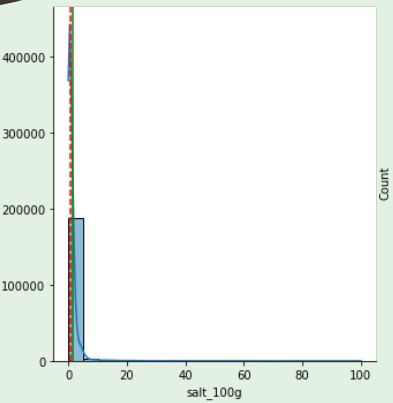
193799 rows × 24 columns

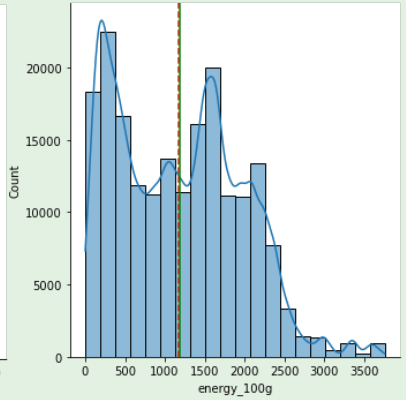
Entrée [ ]:
```
dataKNN2.to_csv("Notebook_openfoodfacts_clean2.csv")
```

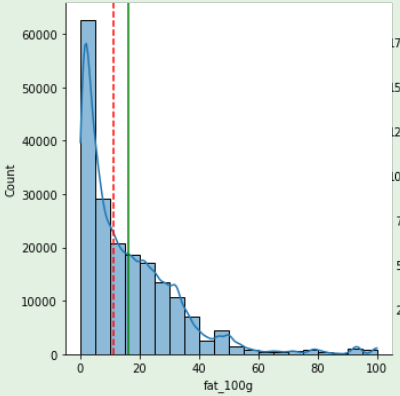
Analyse exploratoire

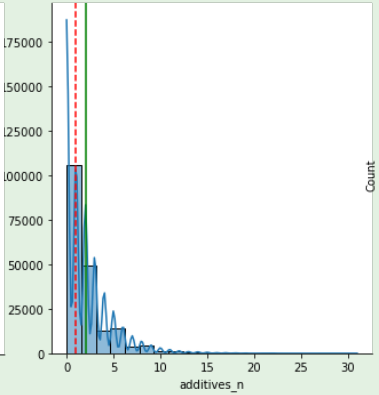


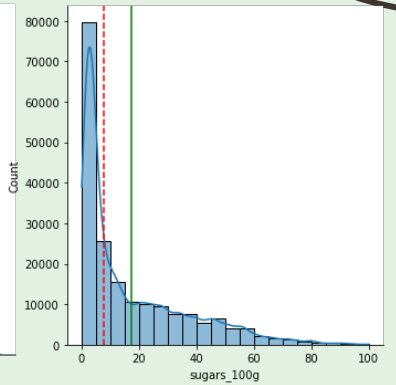
Analyse Univariée

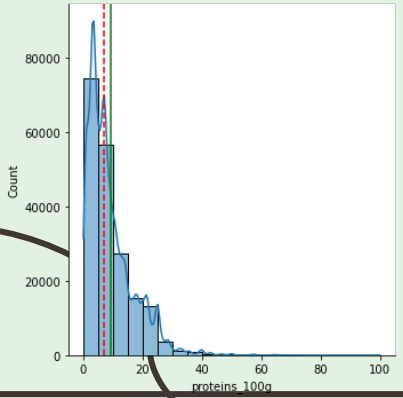


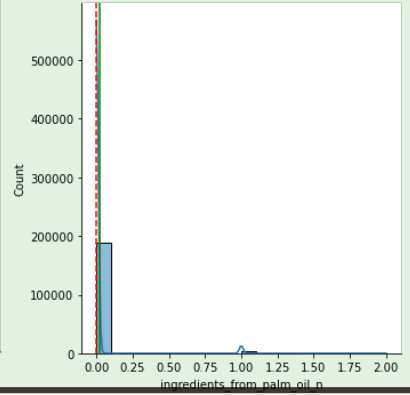


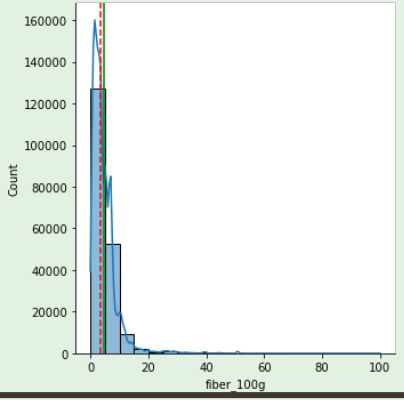










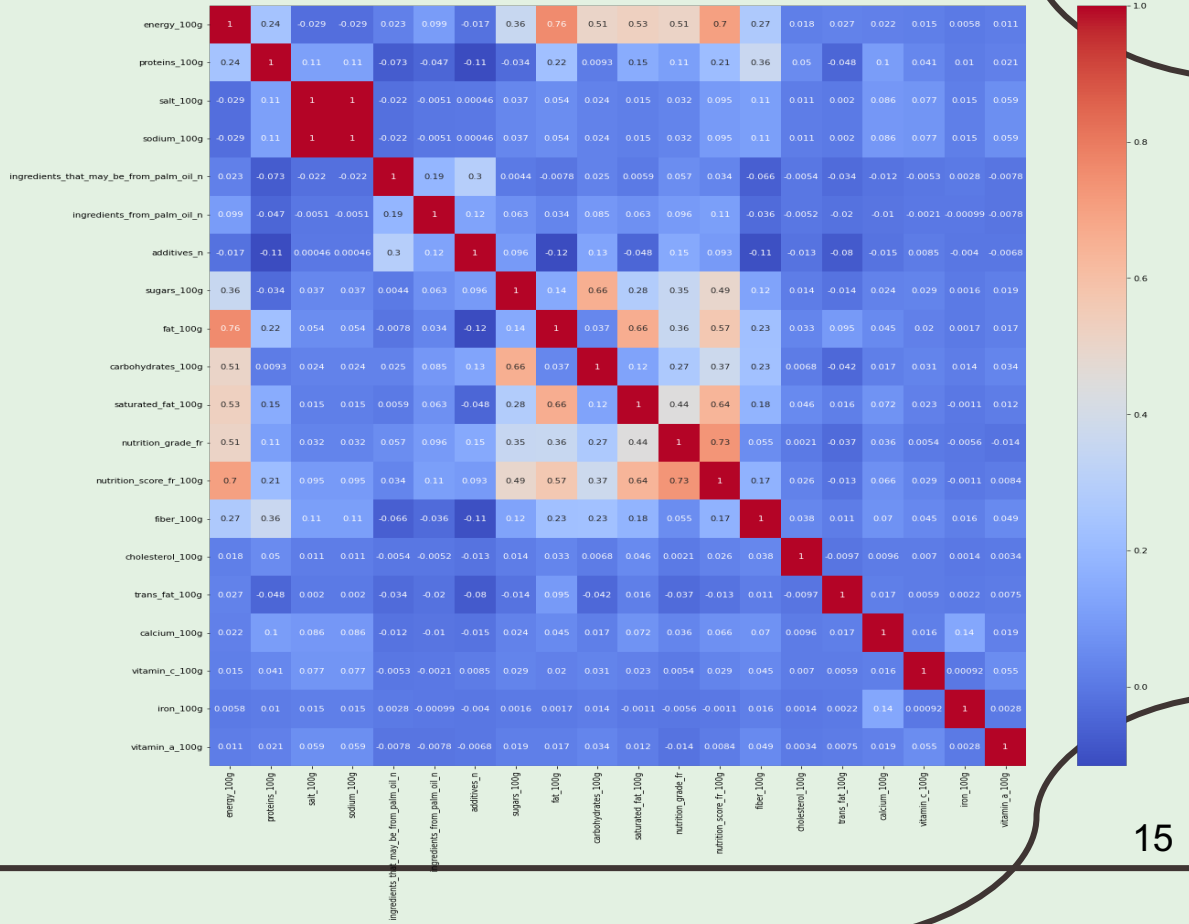


- Pas de loi Normal

Analyse Bivariée

On peut observer des corrélations qui sont logiques entre sucre et carbohydrate, fat et satured fat, sodium et sel.

On observe aussi une corrélation entre le sucre et le carbohydrate avec l'energy.



Loi Normal

Test de Shapiro-Wilk

```
ShapiroResult(statistic=0.644270122051239, pvalue=0.0)
```

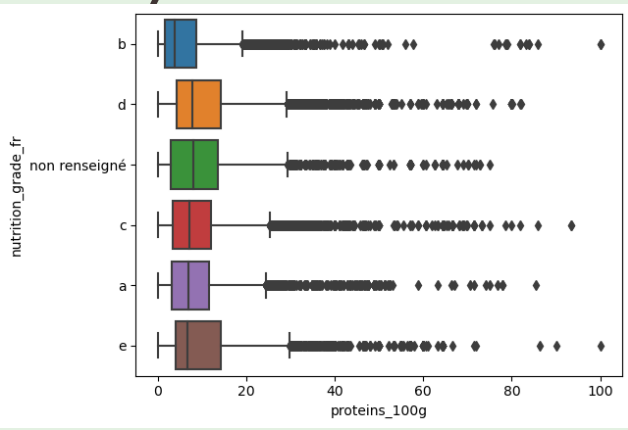
Les données ne suivent pas une loi normale

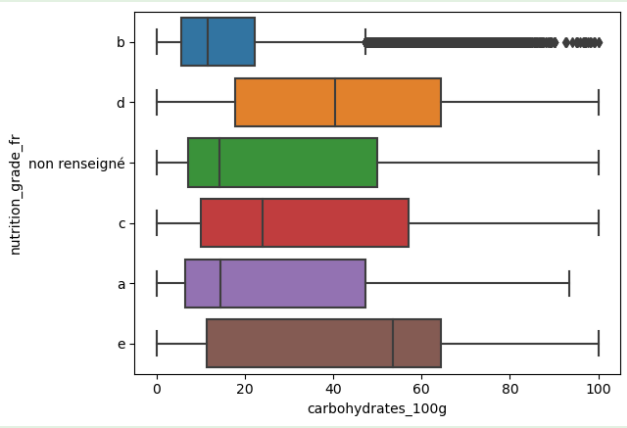
Pas de test Anova

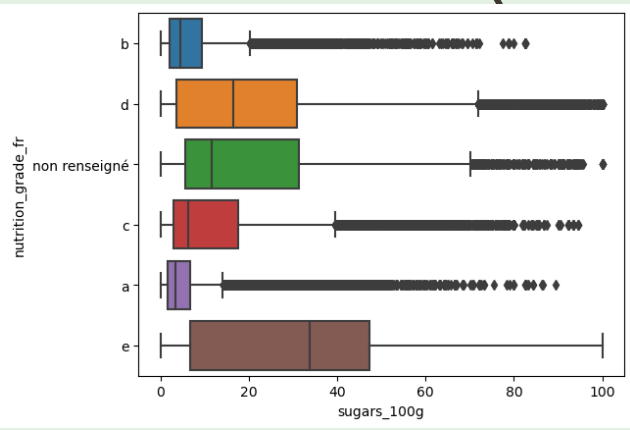
Test de Kruskal Walis

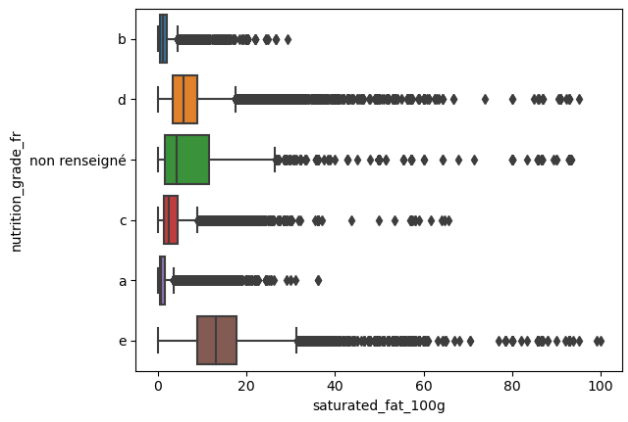
	colonne	statistic	p_value
0	saturated_fat_100g	45883.5	0.0
1	carbohydrates_100g	65704.4	0.0
2	sugars_100g	779.0	0.0
3	proteins_100g	6584.2	0.0
4	salt_100g	217987.3	0.0
5	sodium_100g	271384.6	0.0
6	fiber_100g	70747.4	0.0

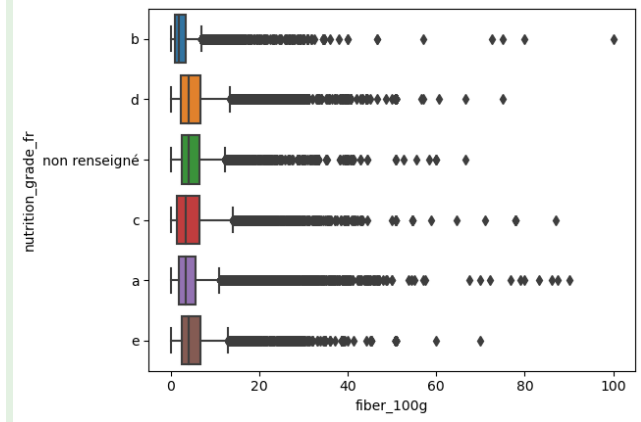
P value = 0, il y a donc bien une dépendance entre les variables et le nutriscore

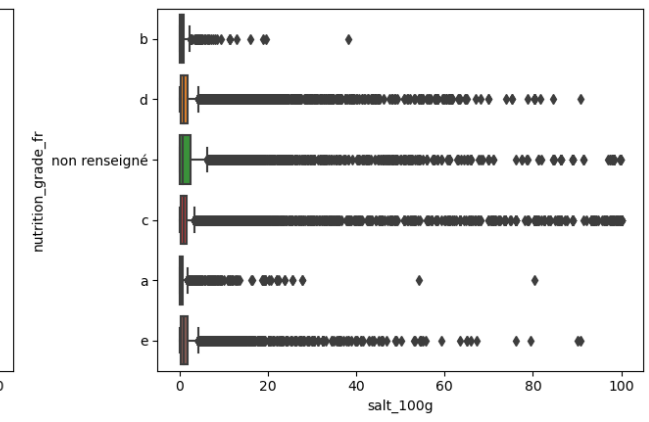

nutrition_grade_fr
b
d
non renseigné
c
a
e
0
20
40
60
80
100
proteins_100g


nutrition_grade_fr
b
d
non renseigné
c
a
e
0
20
40
60
80
100
carbohydrates_100g

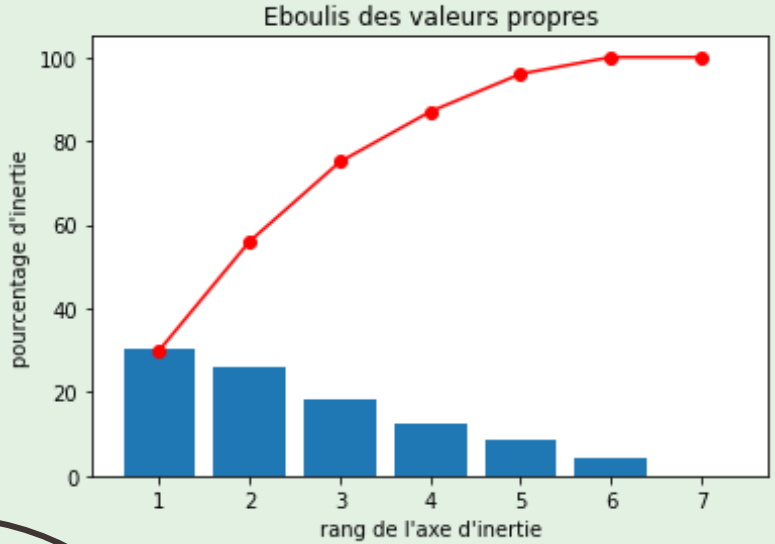

nutrition_grade_fr
b
d
non renseigné
c
a
e
0
20
40
60
80
100
sugars_100g


nutrition_grade_fr
b
d
non renseigné
c
a
e
0
20
40
60
80
100
saturated_fat_100g


nutrition_grade_fr
b
d
non renseigné
c
a
e
0
20
40
60
80
100
fiber_100g


nutrition_grade_fr
b
d
non renseigné
c
a
e
0
20
40
60
80
100
salt_100g

Études de l’Analyse / ACP

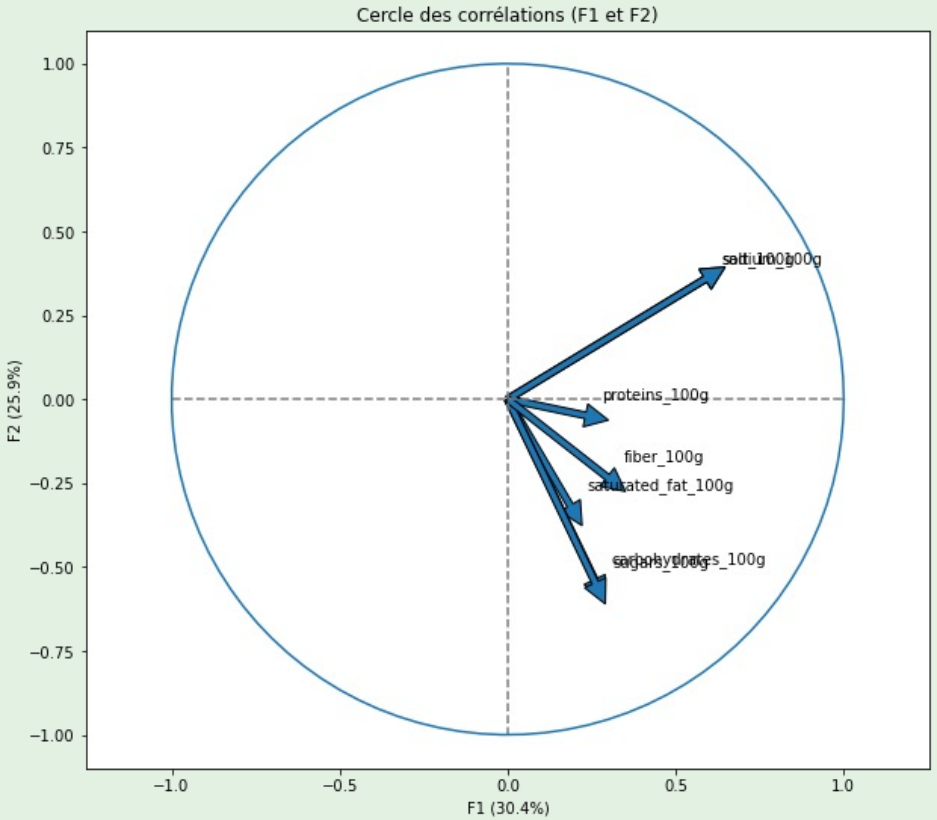


	Composante principale	Pourcentage de variance expliquée	Pourcentage de variance expliquée cumulé
0	1	30.4	30.4
1	2	25.9	56.3
2	3	18.3	74.6
3	4	12.4	87.1
4	5	8.6	95.7
5	6	4.3	100.0
6	7	0.0	100.0

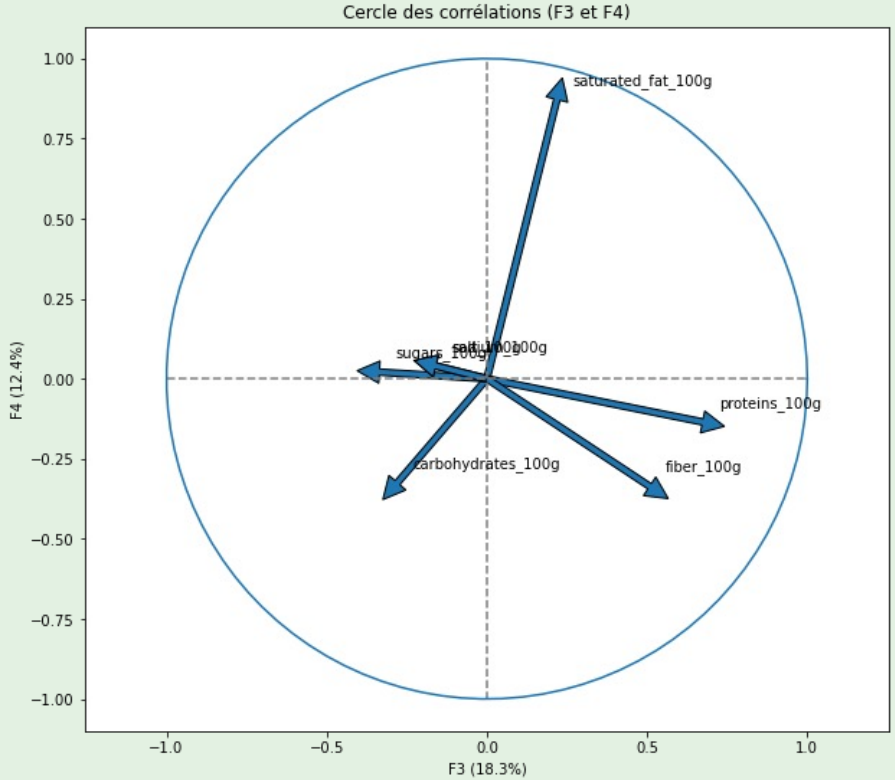
Réduction des données à 87,1% en utilisant 4 composantes principales.

- F1 représente surtout
les produits sucrés, on observe naturellement une corrélation entre le sucre et carbohydrate
- F2 représente les produits salés, on observe naturellement une corrélation entre le sel et le sodium.

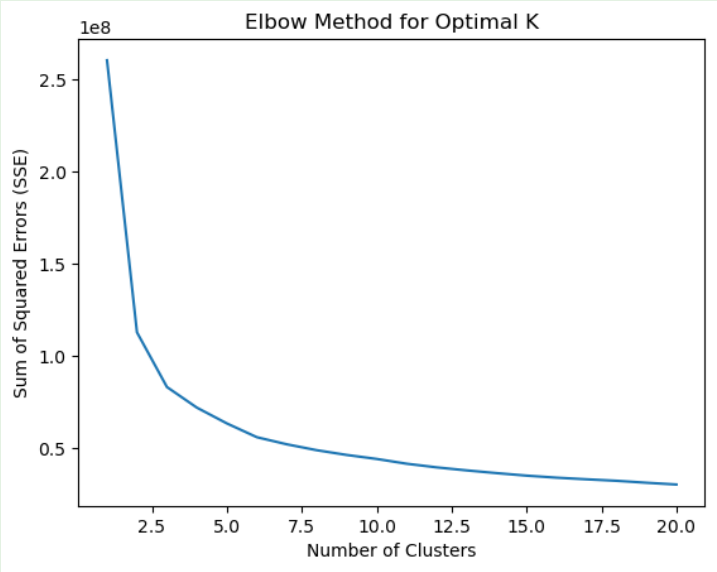
Une petite flèche dans un cercle de corrélation correspond a une faible corrélation.
Il est préférable de n'interpréter que les flèches les plus longues, car les flèches les plus petites correspondent à des variables dites « mal représenté ».



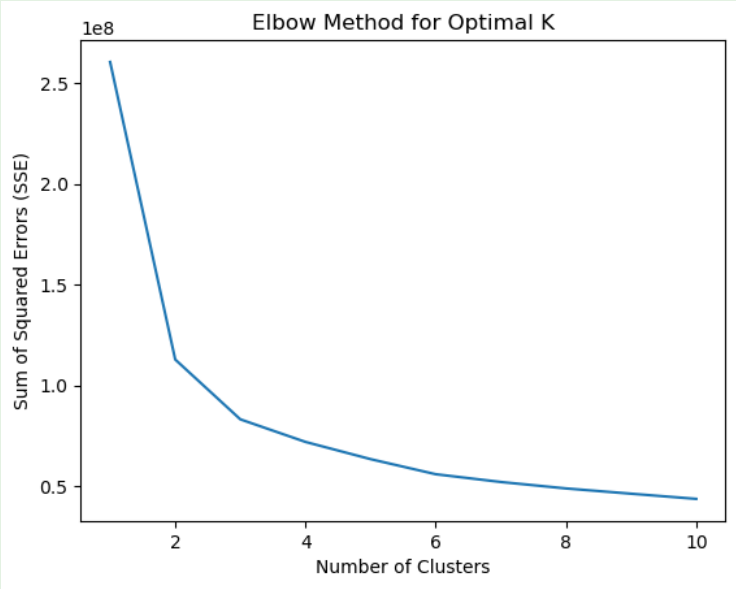
- F3 représente les produits fort en protéine et fibre, une assez bonne corrélation entre eux
- F4 représente les produits forts en gras saturé



K-Means



Une cassure entre 2.5 et 5

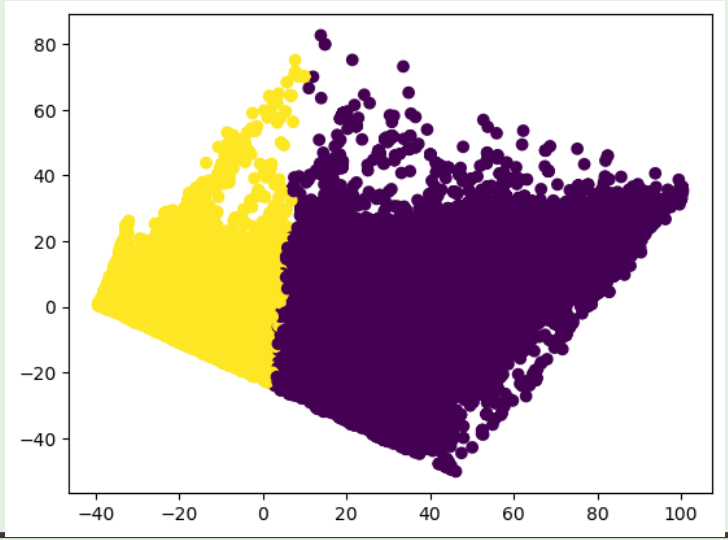


Une cassure entre 2 ou 3

Meilleur cluster

Effectuer une analyse de silhouette

```
Meilleur nombre de clusters: 2
Meilleur score de silhouette: 0.504
```



Cluster 1

	nutrition_score_fr_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	fiber_100g	cluster
count	81627.0	81627.0	81627.0	81627.0	81627.0	81627.0	81627.0	81627.0	81627.0
mean	14.7	8.0	63.2	31.2	9.1	1.7	0.7	5.8	1.0
std	6.8	7.7	14.1	21.3	6.8	5.9	2.3	5.4	0.0
min	1.0	0.0	3.3	0.0	0.0	0.0	0.0	0.0	1.0
25%	10.0	2.3	52.9	10.7	5.0	0.2	0.1	2.9	1.0
50%	14.0	5.4	62.8	30.9	7.1	0.8	0.3	4.0	1.0
75%	20.0	12.1	73.0	46.0	11.1	1.5	0.6	7.1	1.0
max	40.0	90.0	100.0	100.0	100.0	100.0	39.4	100.0	1.0

Cluster 2

	nutrition_score_fr_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	fiber_100g	cluster
count	112172.0	112172.0	112172.0	112172.0	112172.0	112172.0	112172.0	112172.0	112172.0
mean	8.1	5.1	13.4	6.9	9.1	1.3	0.5	3.8	0.0
std	6.8	7.1	10.7	7.4	8.7	3.1	1.2	3.8	0.0
min	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	2.0	0.9	4.7	2.0	2.7	0.2	0.1	1.3	0.0
50%	5.0	2.5	10.5	3.8	6.0	0.8	0.3	2.7	0.0
75%	13.0	6.9	20.0	9.8	14.0	1.5	0.6	5.4	0.0
max	32.0	99.9	48.7	88.9	86.4	100.0	39.4	87.5	0.0

- on observe que les quantités d'ingrédients sont plus élevé dans le cluster 1
- il y a plus de produits dans le cluster 2
- le nutrition_score_fr_100g est plus élever en moyenne dans le cluster 1, les produit ont donc un grade moins bon que le cluster.
- le cluster 1 regroupe les produits très sucrés.

Conclusion

Faisabilité de l’application :

- Corrélation entre diffèrent ingrédient.
- Nutriscore a amélioré.
- Nécessite un expert en nutrition pour mettre en place le score ainsi que les coefficients.
- Cluster difficile.

Axe futur d’amélioration pour l’application :

-Système de recette.
-Amélioration de la base de donnée.
- Création d’un nouveau Nutriscore.