PROJET 2 :

TRAITER ET ANALYSER DES DONNÉES NUTRITIONELLES

SOMMAIRE :

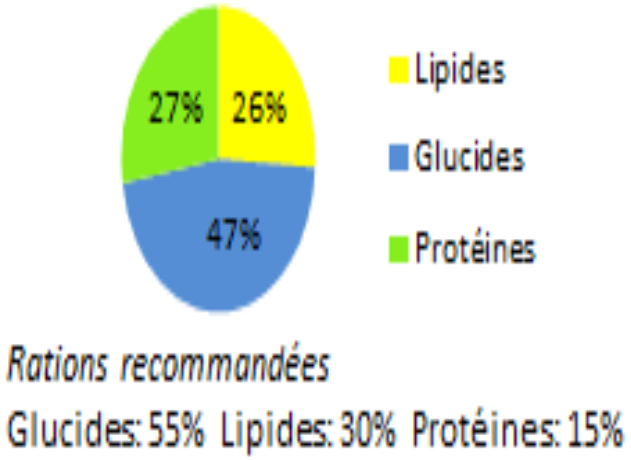
- Présentation du projet
- Objectifs
- Décrire et Nettoyer le DATASET
  - Méthode K plus proches voisins
- Analyse exploratoire
  - -Analyse bivariée
  - Analyse multivariée : Méthode de l'ACP
  - Synthèse
- Conclusion

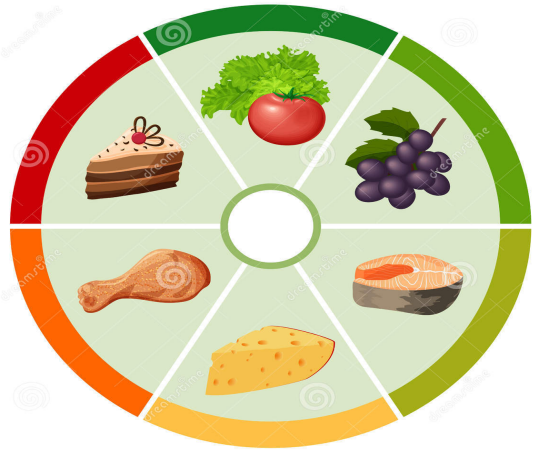
CADRE GLOBAL & OBJECTIFS :

- L'entreprise LAMARMITE souhaite créer un générateur de recette saines
- Réaliser une analyse exploratoire de données nutritionnelles à l'aide d'un DATASET(320 000 lignes, 162 variables)

Recette saine :

Une recette saine est une recette équilibrée et contient moins de glucides et mauvais gras et plus de légumes





1 - DÉCRIRE ET NETTOYER LE DATASET

PREMIÈRE ANALYSE DU DATASET :

- Le DTASET contient 320 000 lignes, 162 variables
- Enormément de valeurs manquantes
- Pour rendre l'analyse possible :

  on garde les variables dont le taux de remplissage > 70%

  on supprime les données inutiles comme La date, l'heure…

DATASET nettoyé :

- 131287 lignes
- 16 variables

DEUXIÈME ÉTAPE : COMMENT TRAITER LES VALEURS MANQUANTES ?

- On supprime les valeurs manquantes tout simplement avec la méthode DROPNA?

  Résultat : perte massive de données

- On utilise l'imputation par la moyenne ?

  Résultat : cette méthode donne une analyse biaisée et n'est pas fiable

ALGORITHME DES K PLUS PROCHES VOISINS

g	trans-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	vitamin-a_100g	vitamin-c_100g	calcium_100g	iron_100g
0	0.000	0.01800	64.29	14.290	3.600	3.57	0.00000	0.000	0.000000	0.021400	0.0000	0.001290
0	0.000	0.00000	60.71	17.860	7.100	17.86	0.63500	0.250	0.000000	0.000000	0.0710	0.001290
0	0.000	0.00000	74.55	25.450	5.500	9.09	0.25400	0.100	0.000000	0.000000	0.0360	0.002620
0	0.000	0.00000	25.00	14.290	7.100	25.00	0.54356	0.214	0.000000	0.000000	0.0710	0.005140

Première étape: entrainer l'algorithme sur une table de référence, sans valeurs manquantes

Deuxième étape : imputer les valeurs manquantes

	energy_100g	fat_100g	saturated-fat_100g	trans-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	vit a
251607	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
183694	945.0	10.00	5.000	NaN	NaN	29.50	10.00	NaN	3.20	0.50000	0.196850	
249618	1142.0	11.00	NaN	NaN	NaN	17.00	NaN	NaN	24.00	NaN	NaN	
247640	736.0	NaN	6.700	NaN	NaN	NaN	16.00	0.70	13.00	0.55000	0.216535	
248251	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

energy_100g	fat_100g	saturated-fat_100g	trans-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	vitamin-a_100g	vitamin-c_100g
1014.5	0.000	0.000	0.000	0.0000	50.000	0.000	0.00	0.00	0.00000	0.000000	0.000000	0.000000
945.0	10.000	5.000	0.000	0.0400	29.500	10.000	1.95	3.20	0.50000	0.196850	0.000095	0.002255
1142.0	11.000	2.650	0.295	0.0735	17.000	1.480	0.90	24.00	1.54686	0.965500	0.000020	0.000600
736.0	4.955	6.700	0.000	0.0240	30.775	16.000	0.70	13.00	0.55000	0.216535	0.000099	0.006350

1 – Choisir le vecteur Target qui contient les « NAN », calculer la distance euclidienne entre les autres points.

2- Calcule de la distance euclidienne entre les points.

$$Dist(c_1, c_2) = \sqrt{\sum_{i=1}^{N} (attr_i(c_1) - attr_i(c_2))^2}$$

$$PlusProcheVoisin = MIN_j(Dist(c_j, c_{test}))$$

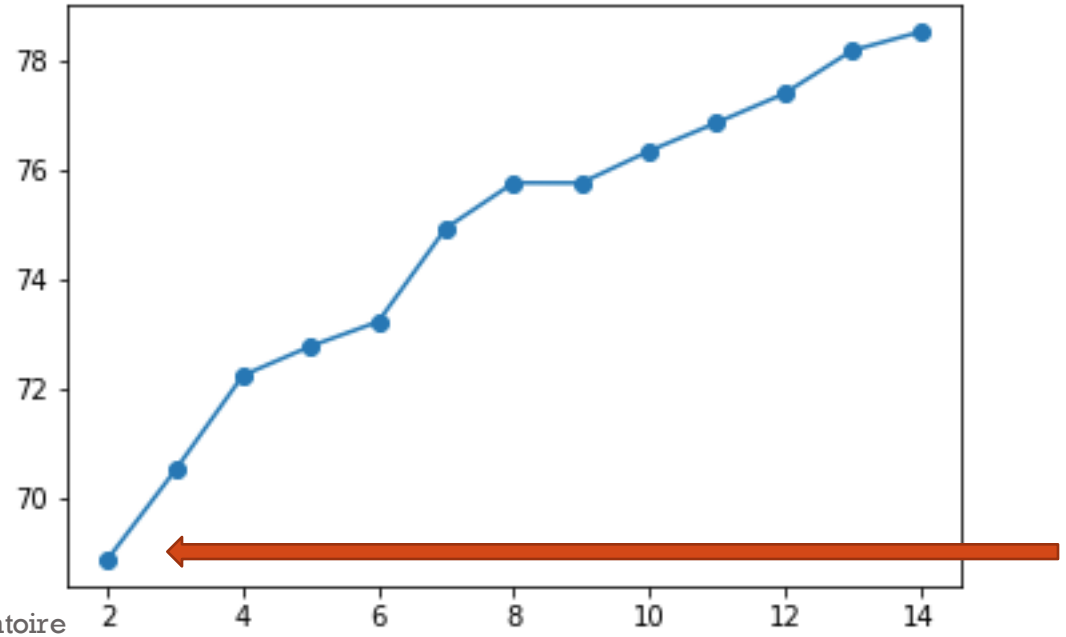
3 - Récupération des K plus proches voisins.

4 - imputer les valeurs manquantes.

K PLUS PROCHES VOISINS – COMMENT TROUVER LE K OPTIMAL ?

- Valeur qui minimise score(k) (erreur quadratique moyenne)

- Création d'un échantillon de 5000 valeurs qu'on s'épare entre test et apprentissage.



K optimal = 2

2 – ANALYSE EXPLORATOIRE

2 – 1 - ANALYSE BIVARIÉE

- Objectif : avoir une première idée sur l'évolution des variables par rapport à la variable illustrative (nutrition grade)

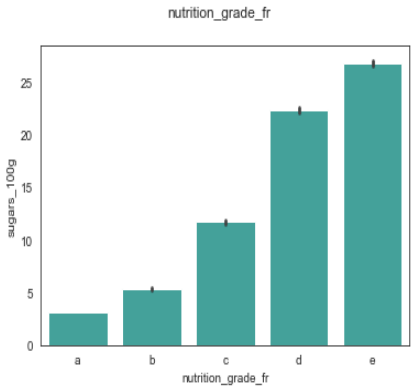


FIGURE 1

On constate que la qualité des aliments baisse quand on a une grande quantité de sucre dans les aliments.

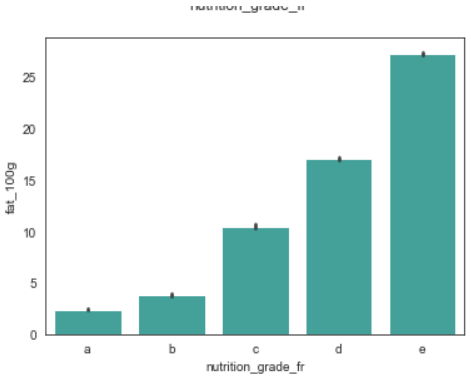


FIGURE 2

Les aliments FAIBLE EN GRAS ont la meilleure qualité.

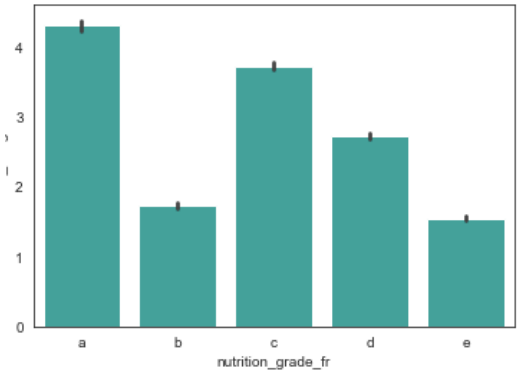


FIGURE 3

La quantité de fibre a une influence sur la qualité des aliments mais pas seulement. Il y a d'autres facteurs.

2 – 2 ANALYSE MULTIVARIÉE : MÉTHODE DE L'ACP

- Objectif réduire le nombre de dimensions tout en conservant un maximum d'informations , pour permettre la représentation graphique (c'est impossible de représenter graphiquement plus de 3 dimensions)
- Réaliser des projections sur des plans en 2 dimensions appelées composantes principales .

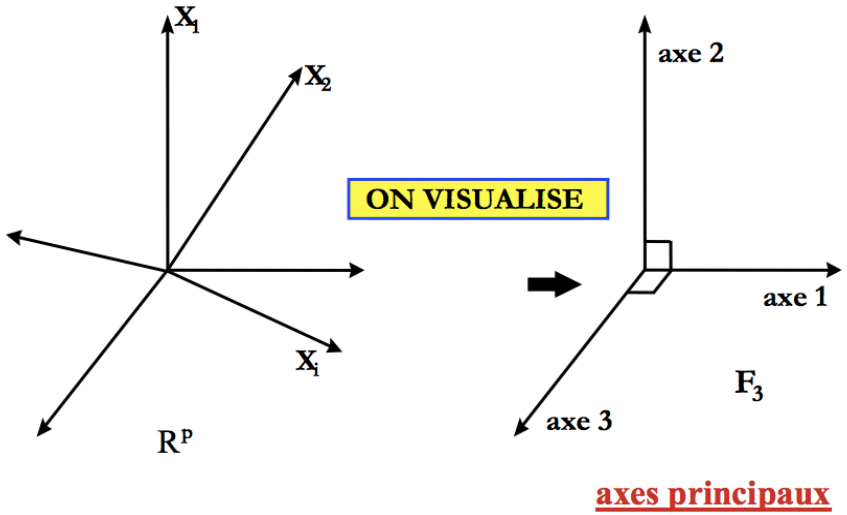
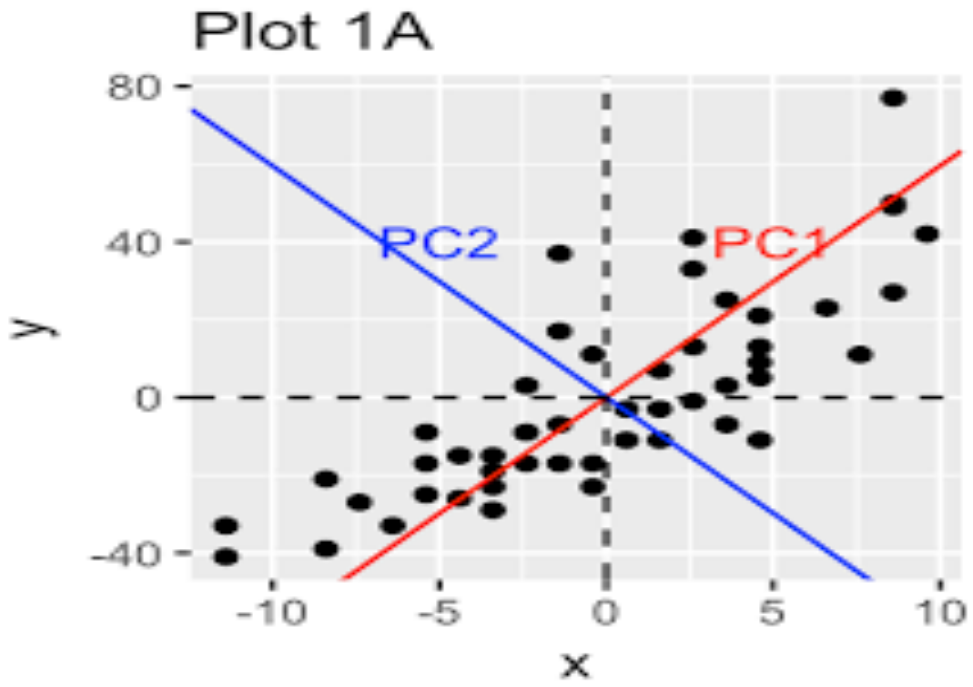


Figure 1

MÉTHODES DE L'ACP :

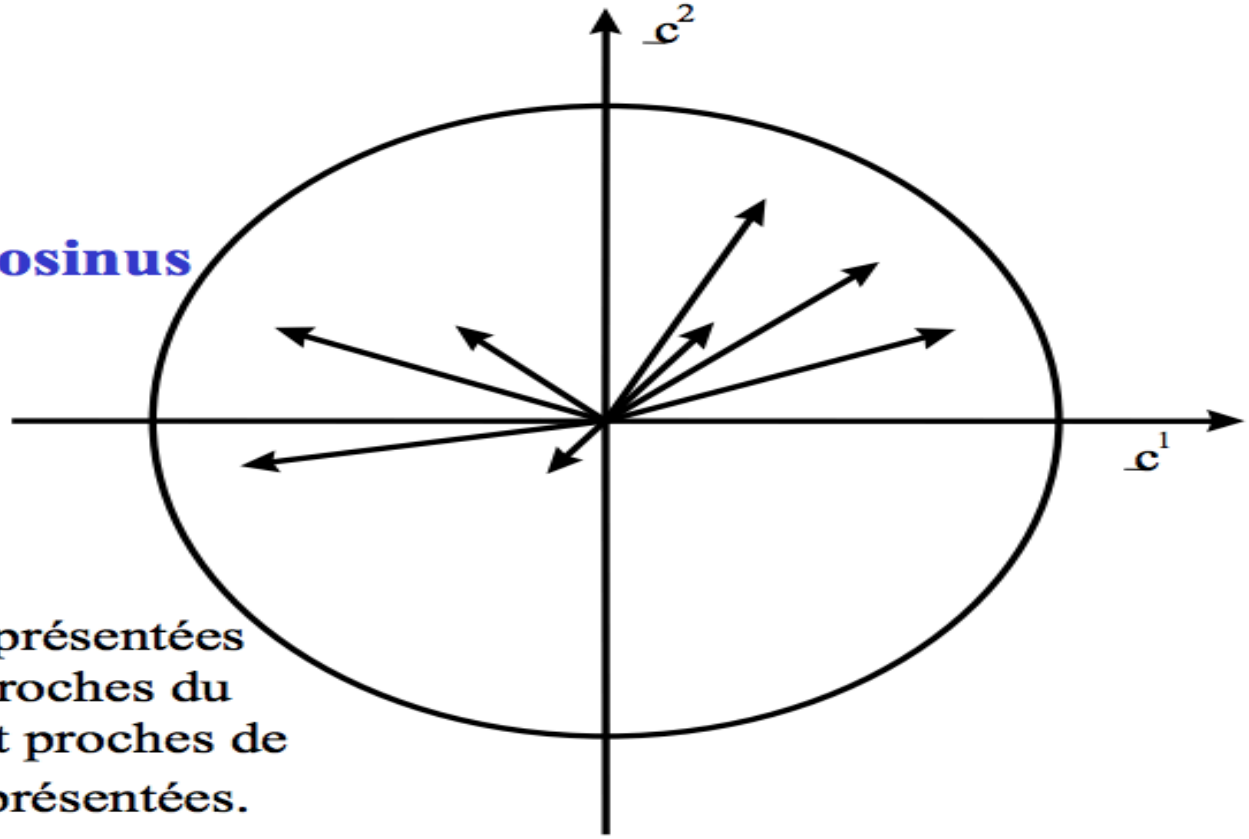
- Les composantes principales sont non corrélées deux à deux. En effet, les axes associés sont orthogonaux.



Figures 2

REPRÉSENTATION DES VARIABLES & INTERPRÉTATION DU CERCLE DE CORRÉLATIONS

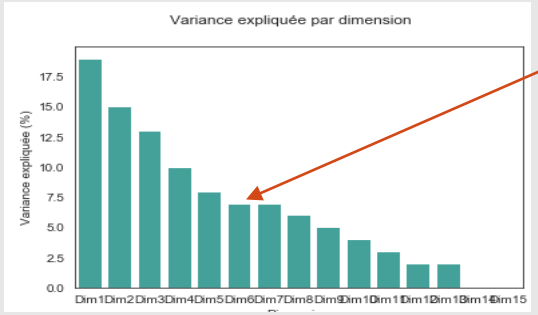
corrélation = cosinus



Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

ACP – CHOISIR LE NOMBRE DE COMPOSANTES

Critère du coude :
Axes avant le point de décrochement



Changement de pente

FIGURES 3

Critère de Kaiser :
Axes associés à des valeurs propres >1

Choix pour ce projet
Nombre de CP est 5, les 5 composantes correspond à 65 % **cum. var. expliquée**)

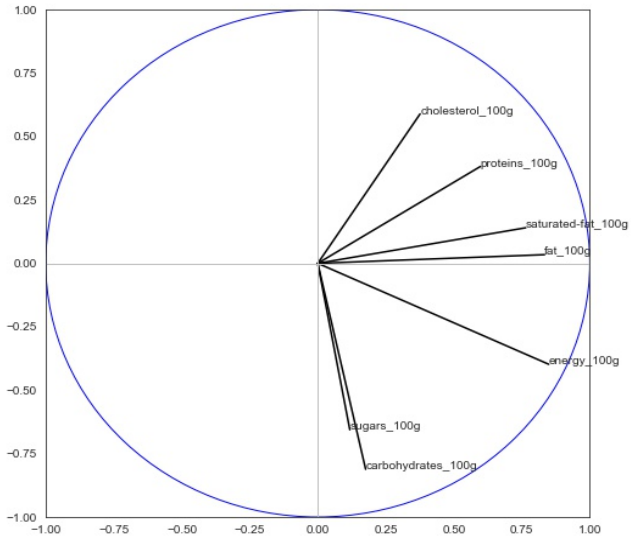
CP = 6(6 valeurs propres >1 et les 6 composantes correspond à 71 % **cum. var. expliquée**)

```
Entrée [250]: #test des bâtons brisés
              print(pd.DataFrame({'Val.Propre':acp.explained_variance_,'Seuils':bs}))
```

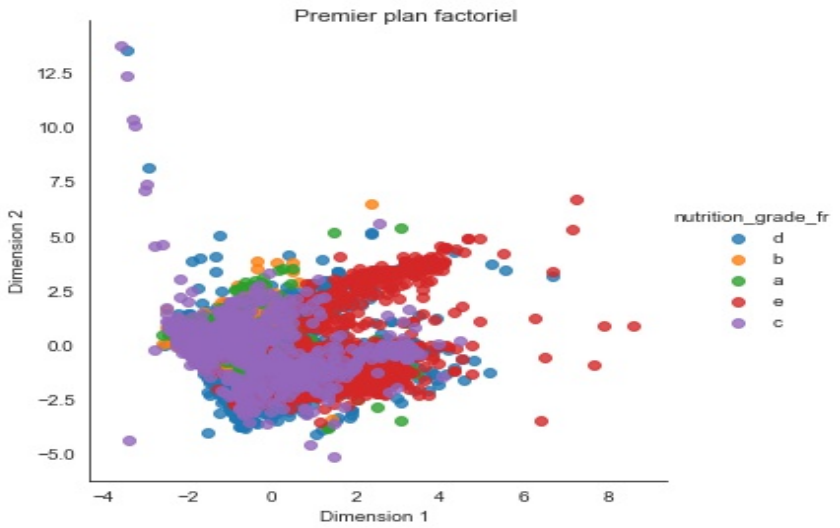
	Val.Propre	Seuils
0	2.976450	3.318229
1	2.431529	2.318229
2	2.007077	1.818229
3	1.562536	1.484896
4	1.200742	1.234896
5	1.018602	1.034896
6	0.914587	0.868229
7	0.725792	0.725372
8	0.645452	0.600372
9	0.569549	0.489261
10	0.407752	0.389261
11	0.298024	0.298352
12	0.236461	0.215018
13	0.008403	0.138095
14	0.000044	0.066667

ACP – ANALYSES MULTIVARIÉES DES ACP 1 ET 2

- Tracer les projections du nuages de points en deux dimensions, ainsi que les cercles de corrélation afin de tirer des conclusions.
- Designer la nutrition grade comme une variable illustrative.



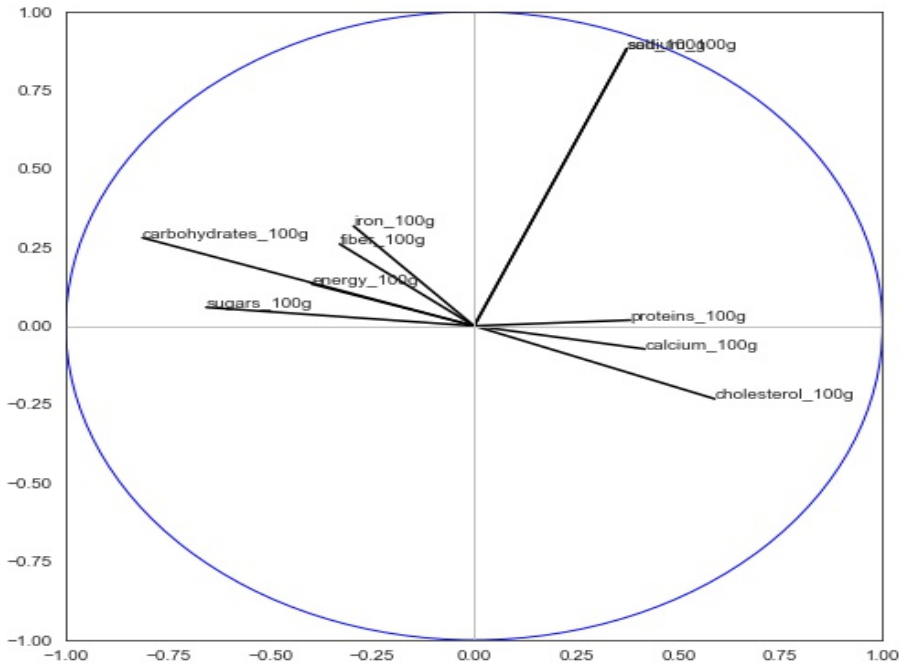
Figures 4



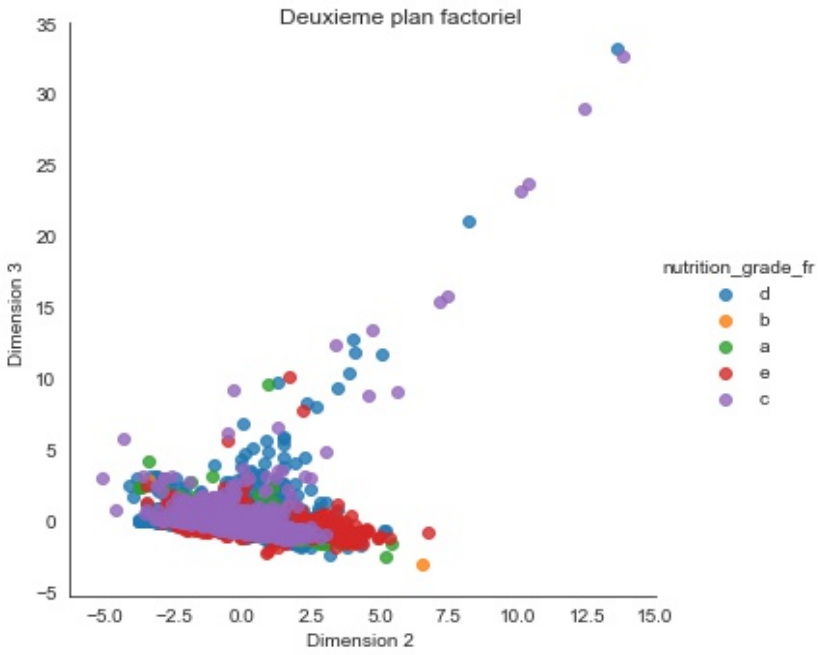
Figures 5

- Sucre et Carbohydrates sont corrélés.
- Sucre et Protéines sont complémentaire (90°).

ACP – ANALYSES MULTIVARIÉES DES ACP 2 ET 3 (DEUXIÈME PLAN FACTORIEL)



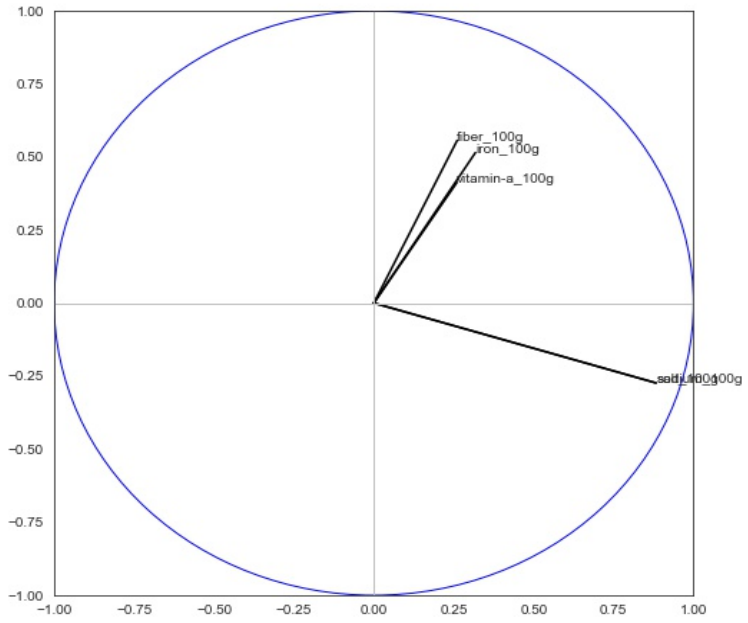
Figures 6



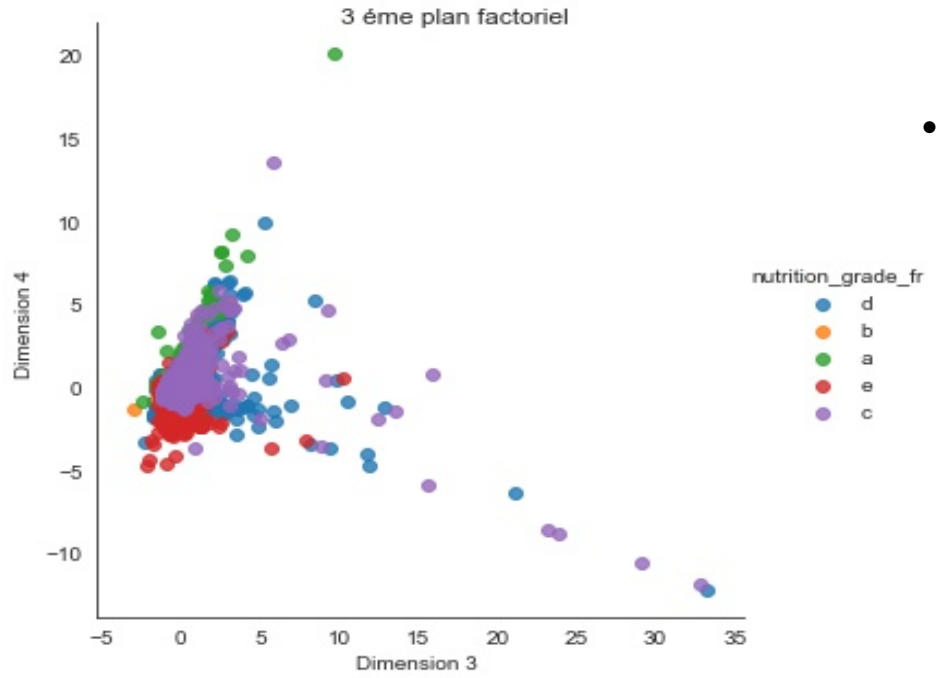
Figures 7

- Cholestérol et Fibre sont corrélés négativement.
- Sucre et Sel sont complémentaire (90°).
- Sel et Energie sont complémentaires.

ACP – ANALYSES MULTIVARIÉES DES ACP 3 ET 4 (3 SÈME PLAN FACTORIEL)



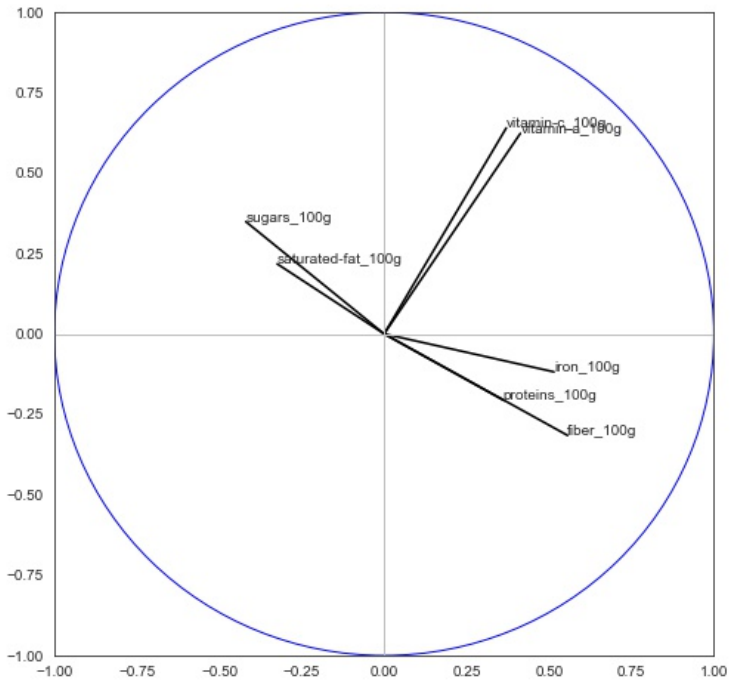
Figures 8

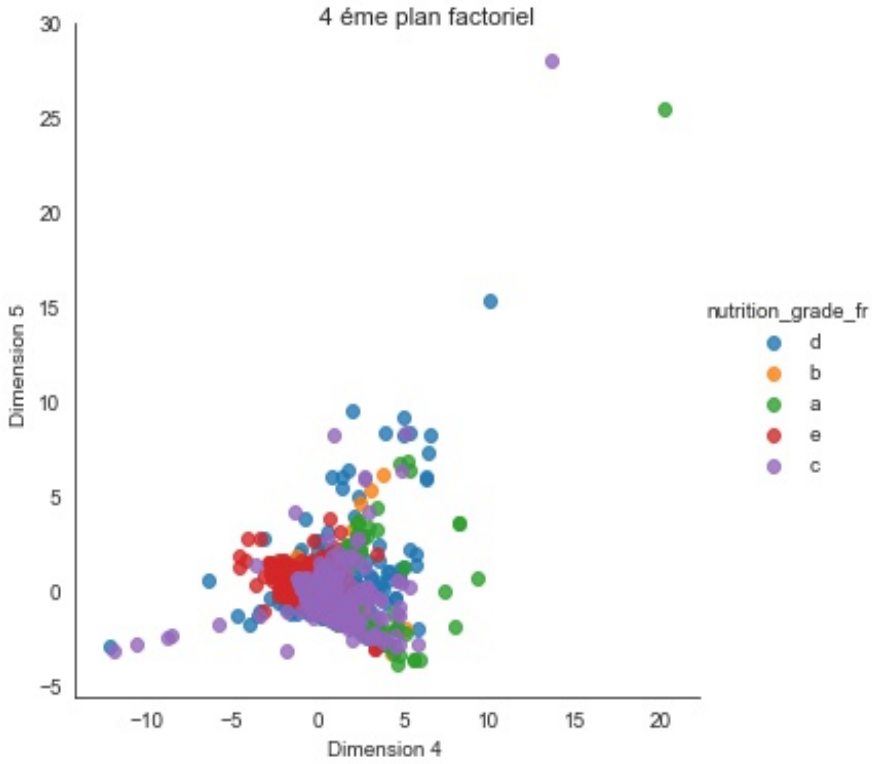


Figures 9

- Fibre et Vitamines et minéraux sont corrélés.
- Fibre et Sel sont complémentaire (90°)

ACP – ANALYSES MULTIVARIÉES DES ACP 4 ET 5 (IVÈME PLAN FACTORIEL)





- Fibre et Protéines sont corrélés négativement avec Sucre .
- Sucre et Satured fat sont corrélés.

SYNTHÈSE DE L'ACP :

NUTRITION GRADES :	
A	La nutrition saines est celle qui contient moins de sel, moins de sucre, moins de matière grasse , riches en protéines et en fibres .
B	Une nutrition saine peut contenir du sucre.
C	La graisse, sucre et les protéines est un vecteur d'énergie contrairement au sel .
D	Les aliments à éviter, sont trop gras trop sucrés , trop énergisants. Nutrition grades faible
E	Les mauvais aliments sont ceux qui contient trop de mauvais gras : Satured fat, Transformed fat …

CONCLUSION :

- La préparation du DATASET est l'étape la plus importante dans une analyse exploratoire (60 % du temps de travail)
- L'utilisation du KNN pour imputer les données manquantes affin d'avoir un résultat cohérant et qui représente parfaitement la data initiale.
- L'ACP est une méthode d'analyse exploratoire qui nous permets de représenter graphiquement les données dans plusieurs dimensions.

REMERCIEMENT :

Je tiens à remercier Monsieur Walid AYADI, mon mentor, pour son aide et ses conseils qui m'ont permis d'avancer efficacement dans ce projet.

Ses connaissances et ses encouragements étaient des éléments essentiels pour accomplir cette mission.