UNIVERSITÉ PARIS DESCARTES

Analyse De Données

Professeur

Nicoleta ROGOVSCHI

Groupe de projet : sujet 5

Mohamed BEN HAMDOUNE
Joseph GESNOUIN
Tony SOREL

Table des matières

Introduction

Présentation

Le jeu de données traité au cours de ce projet porte autour des principaux produits céréaliers produit industriellement qui sont proposés à la vente dans les magasins américains ces dernières années.

L'alimentation et tout particulièrement la "Healthy Food" restent une préoccupation majeure pour beaucoup de personnes de nos jours, certains états encouragent même leur population à mieux consommer mais aussi plus raisonnablement pour réduire les risques de santé liés à une consommation supérieure aux besoins quotidiens moyens recommandés par l'OMS.

Les recherches actuelles indiquent que les adultes ne devraient pas consommer plus de 30% de leurs calories sous forme de graisse, ils auraient également besoin d'environ 50 à 65 g de protéines par jour et devraient fournir le reste de leur apport calorique en consommant des glucides. Ces chiffres donnés à titre d'exemple sont à moduler en fonction du sexe, de l'âge mais aussi de l'activité physique de la personne ciblée.
Un régime recommandé devrait également contenir 20 à 35 grammes de fibres alimentaires.

Le fichier mis à disposition est un résumé des caractéristiques alimentaires de 77 céréales présents dans la majorité des commerces américains:

Les données disponibles comprennent le nom de la céréale, son fabricant, son type de consommation, le nombre de calories par portion, les grammes de protéines, les grammes de gras, les milligrammes de sodium, les grammes de fibres, les grammes de glucides, les milligrammes de potassium, pourcentage typique de l'AJR de vitamines de la FDA, le poids d'une portion, le nombre de « tasse américaine » dans une portion, et l'emplacement de l'étagère.

Une variable nommée "rating" a également été calculée par Consumer Reports, un magazine mensuel américain publié par une association de consommateurs. Cette variable rating permet de classifier la qualité de chaque céréale en fonction des paramètres précisés ci-dessus.

L'objectif principal de cette analyse est de représenter sous forme de graphiques l'essentiel de l'information contenue dans ce jeu de données. Nous allons également essayer d'apporter des éléments de réponse aux questions suivantes: qu'est ce qui qualifie une bonne céréale? Les produits industriels présentés aux consommateurs sont-ils de bonne manufacture comparée aux recommandations journalières des scientifiques? La consommation de céréales est-elle une bonne idée dans le cadre d'un mode de vie sain?

Description des champs du dataset

Pour la plupart des variables définies ci-dessous une visualisation des critères de dispersion est disponible en annexe.

- Le champ « **NAME** » indique le nom des céréales, ici c'est une variable de type chaîne de caractères qualitative, elle nous permet de distinguer chacun des membres du fichier sous l'appellation utilisée en grande surface.

- Le champ « **MANUF** » donne le nom de la compagnie céréalière produisant le céréale (variable qualitative) :
  - A pour American Home Food Products
  - G = General Mills
  - K = Kellogg's
  - N = Nabisco
  - P = Post
  - Q = Quaker Oats
  - R = Ralston Purina

- Le champ « **TYPE** » est une variable booléenne prenant soit la valeur C pour Cold (indiquant froid) et H pour Hot (indiquant chaud).

- Le champ « **CALORIES** » est une valeur numérique de type entier. Pour une barre céréalière suivant sa masse : celle-ci indique le taux de calories présent dans chaque céréale. Dans l'alimentation, les calories sont très importantes à quantifier : elles représentent la quantité d'énergie fournie par la céréale.

- Le champ « **PROTEIN** » indique la masse en gramme de protéine contenue dans une barre céréalière.

- Le champs « **FAT** »: indique la masse en gramme de lipides contenus la barre céréalière (Les Acides Gras saturés compris).

- Le champ « **SODIUM** » indique la masse en milligrammes de sodium (Sel inclus).

- Le champ « **FIBER** » en gramme des fibres alimentaires, valeurs numériques décimales.

- Le champ « **CARBO** » correspond à la masse en gramme des glucides données sous format décimales.

- Le champ « **SUGARS** » correspond à la masse en gramme de sucres (Glucose, Fructose, Lactose,…), c'est une valeur numérique de type entier.

- Le champ « **POTASS** » est la masse en milligramme de potassium sous forme de valeurs numériques entières, le potassium rentre dans les catégories des minéraux. Il est important de noter que c'est le seul minéral décrit dans le fichier.

- Le champ « **VITAMINS** » est un résumé des différents vitamines données sous forme de pourcentage (valeurs entières) suivant les recommandations de la FDA (Food and Drug Administration). C'est une variable qualitative ordinale : trois indicateurs (0%, 25%, 100%).

- Le champ « **SHELF** » prend trois valeurs entières différentes suivant la disposition verticale (basse, moyenne, haute).

- Le champ « **WEIGHT** » est la masse en once (oz) par barre de céréales, à noter que 1 once est égale à 28,34 grammes.

- Le champ « **CUPS** » est le nombre (variable numérique décimale) de « tasse américaine » par barre de céréales, 1 cup est équivalent à 240 millilitres.

- Le champ « **RATING** » est un pourcentage donné sous formes décimales indiquant la notation de la barre de céréales suivant un panel de consommateurs. Cette variable résume la majorité des variables précédemment présentés.

Analyse

Déroulement du processus de traitement

Pour commencer, nous chargeons les données, nous prenons bien en compte que la première ligne contient les noms des colonnes dans ce jeu de données et donc ne doit pas être traitée de la même manière que les autres.

Lors de l'importation des données via la fonction read, on peut constater que le type de données est mis au format de liste, plus précisément un "data.frame".

Sur 3 céréales dans les 77 présents, des valeurs manquantes sont indiquées par "-1.0" ou "-1". Nous avons décidé de remplacer ces valeurs par « NA » afin qu'elle ne soit pas prise en compte pour les statistiques descriptives. Étape fondamentale pour nos futurs traitements et commentaires sur le jeu de données. Sans ce pré-traitement nous n'aurions pas pu avoir des résultats qui puissent être commentés ou nous aurions pu avoir des visualisations biaisées. Nous avons donc décidé d'enlever environ 4% des données pour raisonner sur un jeu de données plus clair

Nous n'avons pas eu énormément de problème durant cette analyse, ceci est dû au fait que la majorité de nos variables sont quantitatives ou qualitatives ordinales et donc faciles à traiter. (Pour plus d'information vous pouvez vous référer à la partie code de l'annexe)

Pendant le traitement de ce jeu de données nous avons utilisés les librairies ggplot2 et corrplot pour construire certains graphiques et dplyr pour faciliter la manipulation de data frame.

Analyse de la présence des compagnies

Nous nous sommes premièrement interrogés sur les producteurs de céréales aux États-Unis, voir ce que nous pouvions en tirer.Le schéma ci-joint représente l'importance de chaque compagnie en fonction du nombre de céréales qu'elle produit.

Nous remarquons (cf Figure 1) qu'il y a deux valeurs extrêmement hautes : General Mills et Kellog's sont plus que majoritaires sur la distribution des céréales aux états unis : à elles seules elles représentent environ 60% des 77 céréales présents dans le jeu de données. Ce sont donc deux grosses compagnies disposant d'une énorme diversité de produits.

A l'inverse, American Home Food Products est présente une seule fois dans l'histogramme ce qui en fait un petit producteur ou du moins avec une diversité de produit limitée.

En dehors de ces trois valeurs extrêmes nous avons donc les producteurs restants: Nabisco, Post, Quaker Oats et Ralston Purina qui représentent à eux tous environ 40% des céréales restants du jeu de données.

Nous remarquons déjà une disparité non négligeable, d'importance et de poids économique des producteurs présentés dans le dataset.

La notation des différentes compagnies

Nous avons donc voulu visualiser la notation moyenne de chacun de ces producteurs en fonction des céréales qu'ils vendaient, voir s'il y avait une quelconque corrélation entre le producteur et la qualité des céréales commercialisées.

Ce graphique (cf Figure 2) nous permet de voir qu'en moyenne la plupart des producteurs se situent au même seuil de notation pour les boites à moustache : Avec une seule valeur il est difficile de conclure quoique ce soit pour le producteur American Home Food Products.

Un seul producteur se distingue positivement de cette comparaison : Nabisco dont les céréales sont notées en moyenne vers 70%.

Il n'y a donc à part pour un seul producteur pas de disparités de qualité visible entre les autres. Y compris pour les deux compagnies majoritaires dont la qualité moyenne des produits ne se distingue pas des autres : la majorité des céréales sont notées entre 30% et 50% à l'exception d'une céréale Kelloggs extrêmement bien notée (90%+): All Bran with Extra Fiber.

En revanche, les céréales de la marque General Mills sont les moins bien notés du jeu de données mais la disparité la plus importante dans la notation des céréales revient à Quaker dont certaines céréales peuvent être considérées comme de bonne manufacture malgré le fait que la boite à moustache varie énormément dans les valeurs basses pour les premiers quartiles.

Nous en concluons donc que la qualité/rating des céréales n'est en moyenne pas fortement liée au producteur qui la fabrique. Il serait cependant intéressant dans le jeu de données d'avoir une variable prix qui nous permettrai d'expliquer cette différence immense de qualité entre les céréales de la marque Nabisco et les autres ou pourquoi les plus petits producteurs ont tendance à produire des céréales de meilleure qualité mais ne disposant pas d'une telle valeur, nous ne nous contenterons simplement de faire des spéculations.

À la vue de cette visualisation nous avons voulu déterminer pourquoi la céréale All Bran with Extra Fiber était si bien notée, et ce qui pouvait la distinguer autant de toutes ses consœurs pour avoir une note aussi phénoménale.

Principales variables dans le jeu de données.

Nous avons réalisé une analyse en composantes principales du jeu de données : cela nous à permit de réduire le nombre de variables et de rendre l'information des données moins redondante : Nous transformons les variables corrélées en axes principaux, chacun décorrélés les uns des autres. Cela nous permet de traiter sur des jeux plus clairs. Mais aussi de remarquer certaines corrélations entre les variables.

Nous pouvons remarquer grâce à cette ACP que :

- Les axes 1 et 2 représentent 52% de l'inertie totale : la première composante principale représente 27.66% de l'inertie totale, la seconde composante représente elle 24.34% de l'inertie totale. Ces deux composantes sont donc relativement importantes dans l'analyse. (cf Figure 3.2)

- Les variables les mieux représentées sur le plan 1-2 sont rating, Potass, Fiber, Weight et calories : celles si sont très proches du cercle de corrélation et donc très bien représentées sur le mapping. (cf figure 3.1)

- Vitamins, sodium et shelf sont très peu représentées et également assez éloignées du cercle de corrélation. Celles-ci sont donc modérément représentées sur le mapping.

- Les variables qui participent le plus à la création de l'axe 1 sont respectivement rating (20.5%), Fiber (19.3%) et Potass (14.7%). Ces trois variables sont plus étroitement liées à cette première composante que les autres.

- Les variables qui participent le plus à la création de l'axe 2 sont respectivement Weight (18.9%), calories (12.9%) et Sugars (11.8%).

On remarque que les deux premiers axes séparent les individus en plus ou moins deux groupes. La séparation s'est majoritairement faite en fonction des variables citées précédemment.

En s'intéressant aux points bien représentés et donc les plus éloignés du centre, on peut remarquer que l'axe 1 représente des céréales tel qu'All bran with ExtraFiber qui sont des céréales avec en moyenne plus de fibres et de potassium que les autres ainsi qu'une note plus qu'honorable. Quant à lui l'axe 2 représente des céréales tel que Puffed Rice ou bien Total Raisin Bran qui quant à eux se caractérisent par un nombre de calories, de sucre et de poids conséquents. (cf Figure 3.3)

Corrélation entre les variables du jeu

Pour la matrice de corrélation, il nous aura fallu changer les valeurs des différents niveaux d'une colonne factor par des valeurs numériques correspondantes afin de réaliser le calcul et exclure la colonne noms inutiles ici pour la corrélation.

Nous remarquons (cf Figure 4) des informations plutôt logiques et rassurantes, tel que le fait que le rating soit corrélé négativement à l'apport calorique, au sucre, au sel contenu, ainsi qu'au poids ou bien à la masse en gramme de lipides contenus. Globalement tout ce dont il faut être attentif lors du choix de la céréale afin de ne pas dépasser les recommandations journalières est donc corrélé à la note.

Le potassium, les fibres sont tous deux corrélés positivement. Ce sont des paramètres qu'il est plus que recommandé de regarder pour avoir une céréale correcte afin d'avoir des apports

de vitamines, et de nutriments importants à l'alimentation : tous ces paramètres étant donc logiquement corrélés positivement avec le rating.

Il est étonnant de noter que Sugar a une corrélation plus forte avec rating qu'avec carbo alors que carbo est un sur-ensemble de Sugar.

On remarque également que les céréales qui se mangent quotidiennement (plus de coupes consommées par jour) ont tendance à être placées dans les premières étagères et donc les variables shelf et cup sont négativement corrélées.

Régression linéaire multivariée

La variable RATING étant la variable clé de l'analyse, celle-ci expliquant à elle seule les différences de qualité entre chaque céréale.Nous avons voulu l'exprimer suivant un modèle linéaire. Cela permettra de pouvoir prévoir ou d'approximer la note d'un produit inconnu lors de son entrée sur le marché, et donc de connaître la qualité de ce nouveau produit.

Nous avons vu que la variable rating était fortement corrélé négativement avec les variables : calories et Sugar.

Dans un graphe présenté en annexe (cf figure 5.1) nous avions essayé de voir s'il nous était possible de modéliser une relation de dépendance entre ces trois variables et le graphe obtenu nous a indiqué que les céréales les mieux notés sont ceux dont l'apport calorique et dont la quantité de sucre sont les plus faibles.

Afin de pouvoir trouver des coefficients viables de prédiction de la qualité des céréales nous avons donc lancé une régression linéaire multivariée (cf Figure 5.2)

Nous remarquons que la probabilité critique(p-value) de se tromper est extrêmement faible (2.2e-16) et donc qu'il est plus que crédible de vouloir modéliser le rating par une relation linéaire entre le sucre et les calories.

De plus le coefficient de détermination est assez élevé (R-squared). La qualité de la prédiction de la régression linéaire est donc assez bonne.

Le modèle généré semble correspondre assez bien à la réalité et peut approximer la note d'une céréale en connaissant ses deux paramètres.

Ce modèle pourrait éventuellement être amélioré en utilisant d'autres variables tel que Fiber ou Potass pour permettre une prédiction plus précise, le résultat de la fonction lm d'un tel modèle est disponible en annexe (cf Figure 5.3) indique un coefficient de détermination bien plus élevé pour une probabilité critique supérieure.

Analyse de l’apport calorique de chaque barre

Il y a précisément 29 barres à 110 calories, 17 à 100 calories et 10 à 120 calories. En moyenne, une barre apporte 106 calories. (cf Figure 6)

Nous pouvons remarquer qu'en moyenne l'apport calorique des barres ne varie pas énormément et que la majorité des barres apportent entre 90 et 120 calories à la personne les consommant, Après avoir effectué un test de Student pour vérifier notre hypothèse, l'apport calorique pourrait se modéliser par une loi normale de moyenne : 106 et d'écart type : 8.8.

Afin d'illustrer au mieux cette information nous avons choisi de les comparer aux AJR dans plusieurs cas de la vie pratique :

Pour une femme, voici un tableau récapitulatif.

Sport\Age	18-40	+40
Aucune activité physique	1900 calories	1750 calories
Active	2150 calories	2000 calories
Grande sportive	2500 calories	2350 calories

On remarque pour une femme avec un tranche d’âge entre 18 et 40 ans, il existe trois cas différents en pourcentage d’apport caloriques en moyenne pour une collation recommandée qui sont 5.57%, 4.93%, 4.24%.
Dans le cas d’une femme de plus de 40 ans, les chiffres sont : 6.05%, 5.3%, et 4.51%.

Pour un homme, le tableau ci-dessous des apports en fonction de l’âge.

Sport\Age	18-40	+40
Aucune activité physique	2350 calories	2200 calories
Active	2650 calories	2450 calories
Grande sportive	3250 calories	3050 calories

Dans une tranche d’âge comprises entre 18 et 40 ans, nous avons en pourcentage : 4.51%,4%,3.26%.
Dans le cas des hommes âgés de 40 ans et plus : 4.81%,4.32%,3.47%.

On en conclut que sur l'échantillon des barres céréalières sur le marché américain dont on dispose, celles-ci peuvent être considérées comme des collations correctes à la vue des apports caloriques moyens apportés par une portion. Il est cependant important de noter que l'analyse réalisée porte seulement sur les calories apportées et non sur les macronutriments (Glucides, Protéines, Lipides).
Lien pour la source d'information sur les calories.

Analyse de l'apport en vitamines des céréales

Afin de compléter ce qui a été précédemment à propos de l'apport calorique considéré comme correct dans le cas d'une collation, nous avons voulu regarder les autres apports significatifs d'une barre de céréale. (cf Figure 7)

Dans le cas présent, nous remarquons que 8 des céréales présentées n'ont aucun apport notable en vitamines, que la plus grosse majorité en apporte un peu et certaines céréales permettent un apport complet des vitamines journalières en une seule consommation.

Cela nous permet de compléter au moins visuellement ce qui a été dit à propos de la barre de céréale en tant que collation en fonction des calories : les barres de céréales permettent en dehors d'apporter de l'énergie, un apport en vitamines plus ou moins important en fonction de la céréale, ce qui n'est pas négligeable dans le cadre d'un mode de vie sain.

Conclusion

Dans ce rapport, nous avons traité un jeu de données contenant différentes céréales et nous avons essayé de modéliser graphiquement un maximum d'informations contenues dans les données fournies.

Le jeu de données n'aura pas nécessité énormément de pré-traitement de par la nature de ses variables facilement manipulable. De plus le fait d'avoir accès à la variable rating nous a fortement aidé dans la modélisation et la classification des céréales fournis.

Nous avons vu de quoi se composait une bonne céréale et nous avons également défini un modèle capable de les identifier. Nous avons brièvement regardé comment se répartissaient les produits proposés au consommateur en fonction des entreprises les produisant et nous en sommes arrivés à la conclusion que les céréales peuvent être une option viable dans le cadre d'une consommation équilibrée suivant les recommandations journalières.
La consommation de céréales est donc un choix adapté dans le cadre d'un mode de vie sain, pour autant comme toute chose, il est nécessaire de bien regarder les caractéristiques de la céréale envisagée et de privilégier une céréale riche en fibre et en potassium et faible en sucre.

ANNEXE

Annexe 1

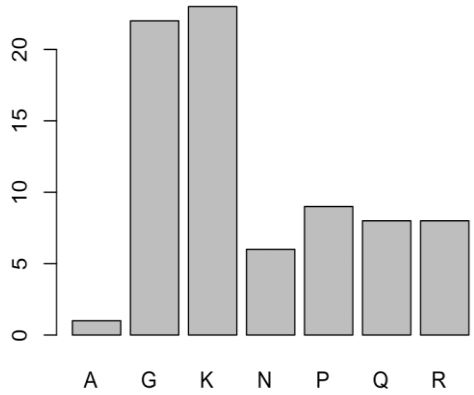


Figure 1 - Répartition des barres de céréales en fonction de leur producteur.

Annexe 2

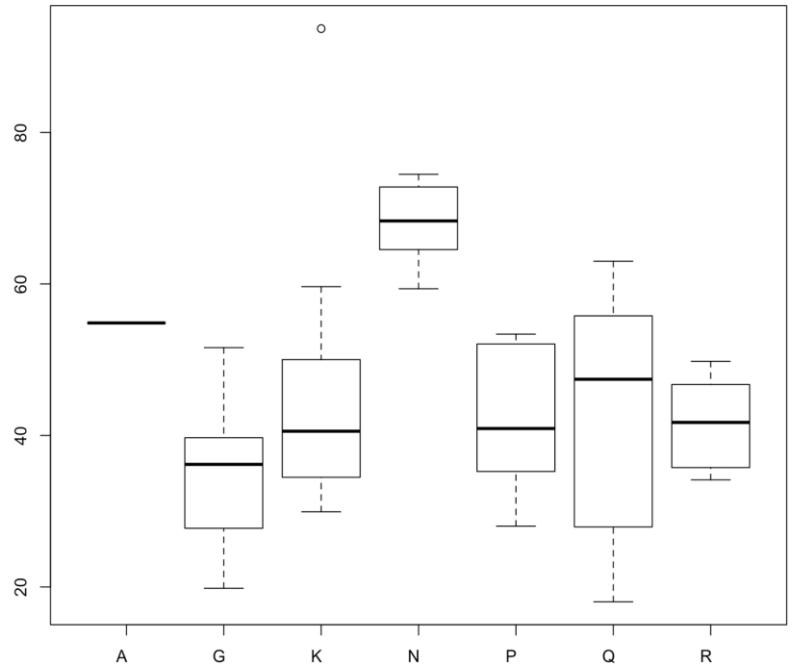


Figure 2 – Diagramme en boîte à moustache des RATING de chaque producteur

Annexe 3

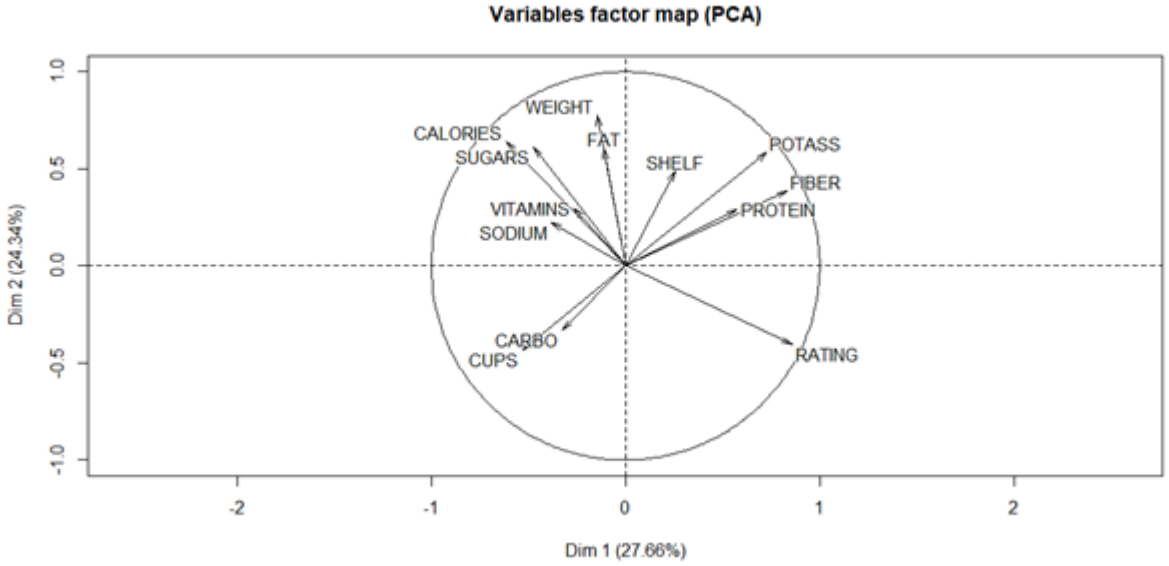


Figure 3.1 – Cercle de corrélations des variables du dataset

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.595614009	27.65856930	27.65857
comp 2	3.164110611	24.33931239	51.99788
comp 3	1.866592981	14.35840755	66.35629
comp 4	1.090914951	8.39165347	74.74794
comp 5	0.969609779	7.45853676	82.20648
comp 6	0.723307242	5.56390186	87.77038
comp 7	0.663425135	5.10327027	92.87365
comp 8	0.438210393	3.37084918	96.24450
comp 9	0.301944077	2.32264675	98.56715
comp 10	0.091659943	0.70507649	99.27222
comp 11	0.065339045	0.50260804	99.77483
comp 12	0.025519139	0.19630107	99.97113
comp 13	0.003752693	0.02886687	100.00000

Figure 3.2 – Valeur Propre de chaque axe principal

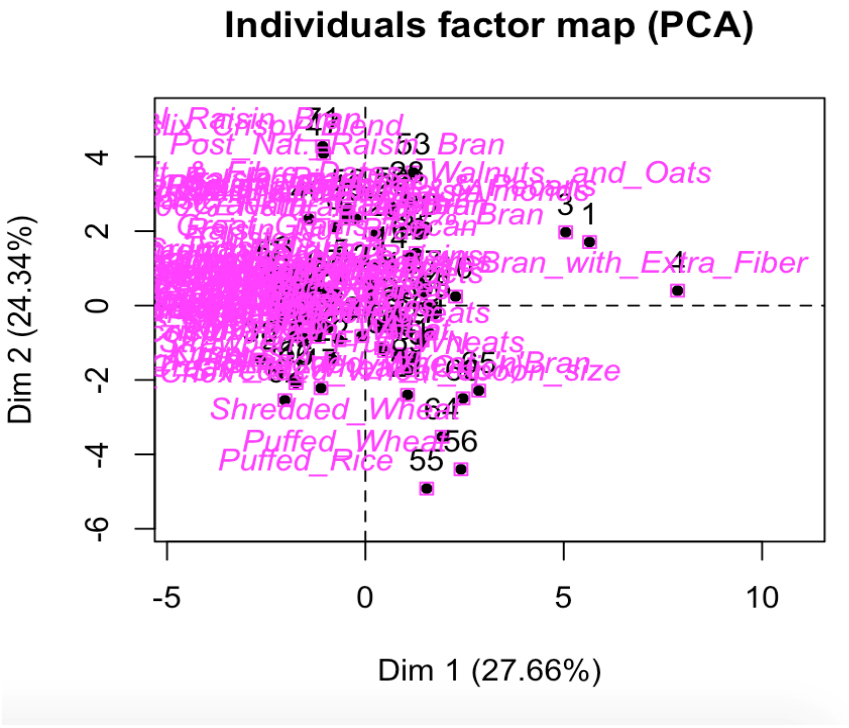


Figure 3.3 - Nuage des individus

Annexe 4

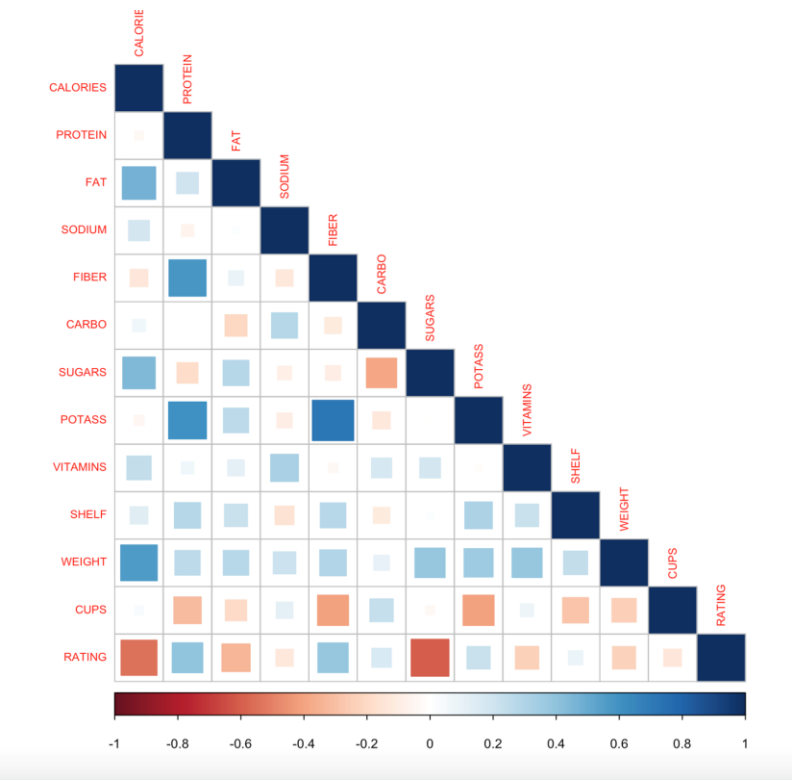


Figure 4- Matrice de corrélation

Annexe 5

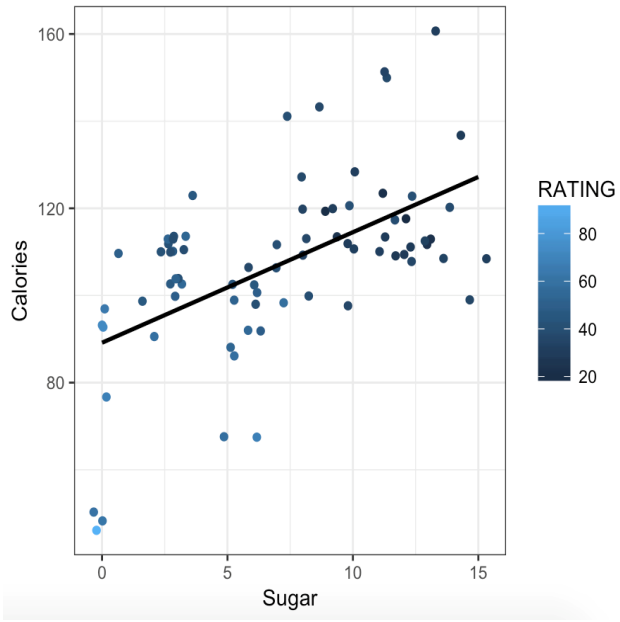


Figure 5.1 Régression linéaire multivariée

```
Call:
lm(formula = RATING ~ CALORIES * SUGARS)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3573  -5.3315   0.1643   4.7627  15.7846

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     101.584447   7.561683  13.434  < 2e-16 ***
CALORIES         -0.449421   0.078401  -5.732 2.16e-07 ***
SUGARS           -5.090190   1.076903  -4.727 1.10e-05 ***
CALORIES:SUGARS   0.031034   0.009813   3.162  0.00229 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.632 on 72 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7191,    Adjusted R-squared:  0.7074
F-statistic: 61.43 on 3 and 72 DF,  p-value: < 2.2e-16
```

Figure 5.2 - Résultat de la fonction lm

```
Call:
lm(formula = RATING ~ CALORIES * SUGARS * POTASS * FIBER)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4593 -2.0831 -0.1437  2.2168 11.3711

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.547e+01  9.023e+00   8.364 1.51e-11 ***
CALORIES                   -3.581e-01  9.803e-02  -3.653 0.000559 ***
SUGARS                      6.411e+00  2.268e+00   2.826 0.006453 **
POTASS                     -2.129e-01  2.735e-01  -0.778 0.439458
FIBER                       1.881e+01  1.185e+01   1.588 0.117670
CALORIES:SUGARS            -6.192e-02  2.074e-02  -2.985 0.004142 **
CALORIES:POTASS             3.944e-03  2.713e-03   1.454 0.151432
SUGARS:POTASS              -4.216e-02  3.021e-02  -1.395 0.168186
CALORIES:FIBER             -1.378e-01  1.194e-01  -1.154 0.253063
SUGARS:FIBER               -3.903e+00  1.829e+00  -2.135 0.037034 *
POTASS:FIBER               -1.458e-02  2.746e-02  -0.531 0.597599
CALORIES:SUGARS:POTASS      1.271e-04  2.670e-04   0.476 0.635721
CALORIES:SUGARS:FIBER       3.688e-02  1.689e-02   2.184 0.033034 *
CALORIES:POTASS:FIBER      -2.521e-04  3.930e-04  -0.642 0.523636
SUGARS:POTASS:FIBER         1.319e-02  4.551e-03   2.899 0.005280 **
CALORIES:SUGARS:POTASS:FIBER -8.610e-05  4.418e-05  -1.949 0.056183 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.322 on 58 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.9247,    Adjusted R-squared:  0.9052
F-statistic: 47.45 on 15 and 58 DF,  p-value: < 2.2e-16
```

Figure 5.3 - Résultat de la régression linéaire en prenant plus de variables en compte

Annexe 6

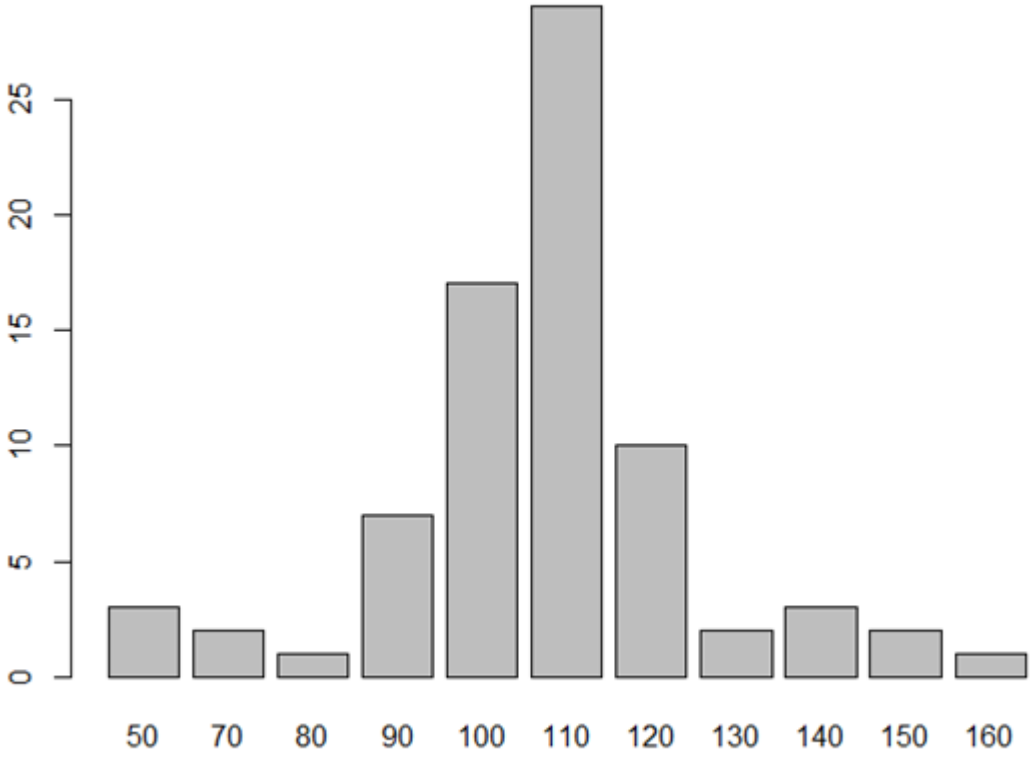


Figure 6
Répartition des CALORIES

Annexe 7

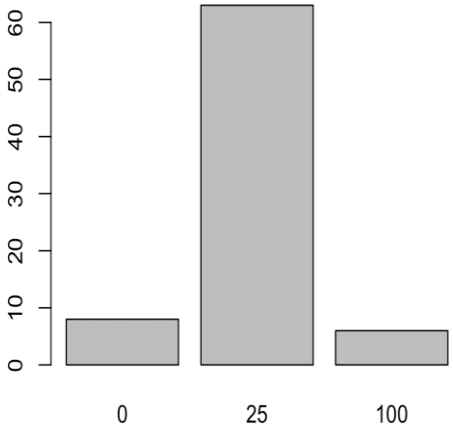
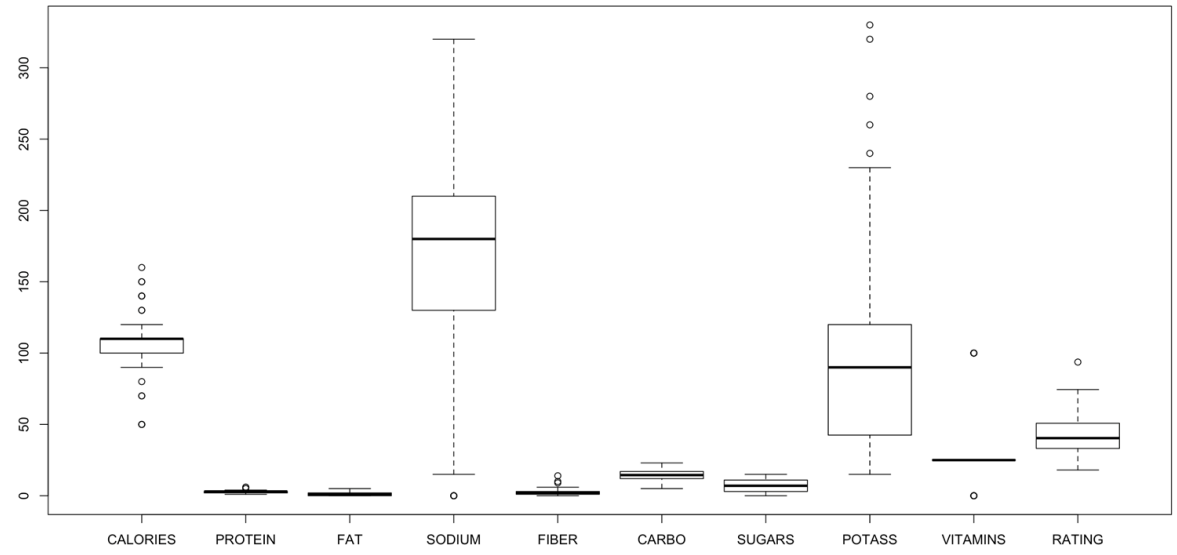
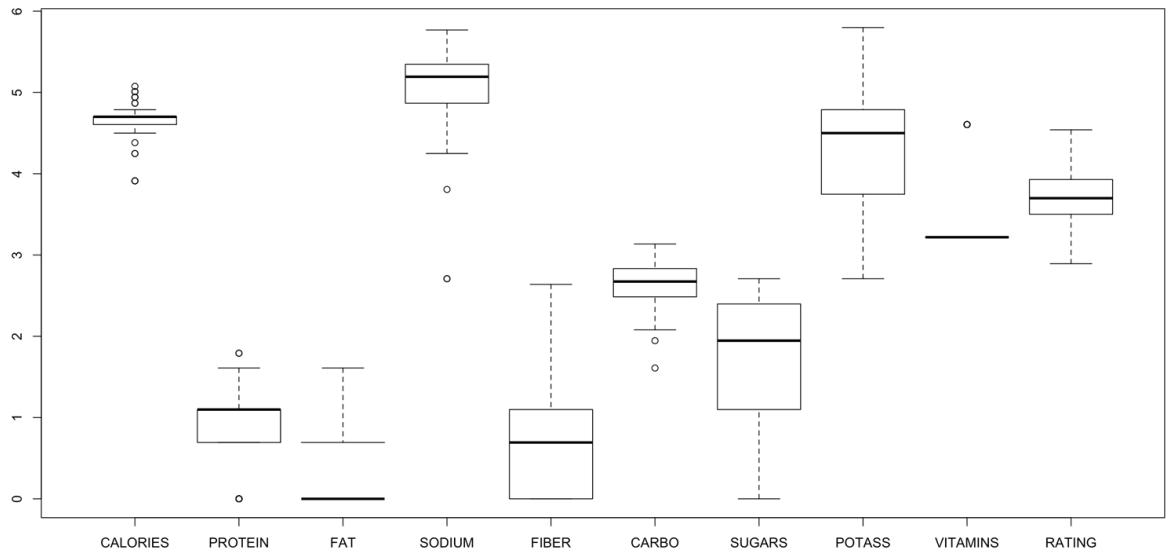


Figure 7 – Répartition des barres en fonction des VITAMINS :
Le champ « VITAMINS » est un résumé des différents vitamines données sous forme de pourcentage (valeurs entières) comprenant trois indicateurs (0%, 25%, 100%) suivant les recommandations de la FDA (Food and Drug Administration).

Annexe 8



Répartition des variables sans log.



Répartition des variables en log.

Code

```
#Importer la librairie FactoMiner
library(FactoMineR)
#Importer la librairie ggplot2
library(ggplot2)
#Importer la librairie corrplot
library(corrplot)
#Importer la librairie dplyr
library(dplyr)


#On charge le fichier
cereal=read.delim("/Users/jzk/Downloads/BD_Projet_2017_2018_INFO/Binome5/276-
cereals.txt",header = T)

#Cette instruction nous permet d'avoir le nom des colonnes contenu dans notre liste
colnames(cereal)

# Cette instruction nous donne le nombre de ligne unique pour ce jeu données, c'est à dire le
nombre de barre céréalières différentes.
nrow(unique(cereal))

#Le nombre de variables dans ce jeu de données
ncol(cereal)

#On souhaite récupérer le type de chaque variable
#On constate que nom est "integer" etc...
sapply(cereal,typeof)

#Récupérer les informations sur les variables et discuter
sapply(cereal,class)

#On remplace les valeurs manquantes par NA afin de ne pas être inclus dans les calculs
cereal[cereal==-1 | cereal==-1.0]=NA

#Aperçu des 10 premières lignes
head(cereal,10)


# ACP
resultatPCA = PCA(cereal,scale.unit = T,graph = T,quali.sup = 1:3)
resultatPCA$eig
resultatPCA$var$coord
resultatPCA$var$contrib

#Matrice de corrélation
numeric_var <- names(cereal)[which(sapply(cereal, is.numeric))]
cor(cereal[,numeric_var], use="complete.obs", method="kendall") %>%
  corrplot(method="square", type="lower", tl.cex= 0.7)
```

```
#On utilise cette fonction attach afin de diminuer la longueur des variables au cours du code.
attach(cereal)

# Un histogramme des plusieurs compagnies présentes, on constate que G,K sont très présents et que A très peu.
table(MANUF)
barplot(table(MANUF))

#Faire une régression linéaire entre rating calories, sugar : lm1=lm(rating~cal*sug)
gg1=ggplot(data=cereal,aes(x=SUGARS,y=CALORIES,col=RATING))+
  geom_jitter(data=cereal,aes(,CALORIES,col=RATING))+
  labs(x="Sugar",y="Calories")+
  geom_smooth(method="lm",se=FALSE,col='black')+
  theme_bw()
gg1

lm1=lm(RATING~CALORIES*SUGARS)
summary(lm1)

###Régression linéaire plus précise
lm2=lm(RATING~CALORIES*SUGARS*POTASS*FIBER)
summary(lm2)
#On enlève les noms variable inutile pour la corrélation
utile=cereal
head(utile)
#On enlève la colonne NOM
utile=utile[,-1]
#Conversion des LEVELS MUNUFACTEUR
utile$MANUF=as.numeric(utile$MANUF)
#Conversion des LEVELS TYPE C OU H
utile$TYPE=as.numeric(utile$TYPE)
#Matrice de corrélation entre les variables
MatrixCor=cor(utile,use = "na")

#Faire un rating par compagnie pour expliquer une influence.
#Boxplot de toutes les compagnies
plot(MANUF,RATING)


#Histogramme entre nombre de barre de céréales et calories (fréquences)
# 110 calories pour une barre
barplot(table(CALORIES))
###Test de Student pour vérifier que cela peut se modéliser comme une loi normale
t.test(CALORIES, mu=106)
```

```
#Décrire pour chaque colonne la valeur la plus basse, la plus haute ainsi que médiane/moyenne.
summary(CALORIES); boxplot(CALORIES, main="Répartition des calories") #Description des calories
summary(RATING); boxplot(RATING, main="Répartition des notes") #Description des votes
summary(SUGARS); boxplot(SUGARS, main="Répartition du sucre") #Description du taux de sucres dans les barres céréalières, enlever valeurs -1
summary(CARBO); boxplot(CARBO, main="Répartition des glucides") #Description des glucides, enlever valeurs -1
summary(FAT); boxplot(FAT, main="Répartition des lipides") #Description des lipides.
summary(PROTEIN); boxplot(PROTEIN, main="Répartition des proteines") #Description des protéines.
summary(SODIUM); boxplot(SODIUM, main="Répartition de la quantité de sel")
#Description de sodium
summary(FIBER); boxplot(FIBER, main="Répartition des fibres") #Description de fiber.
summary(POTASS); boxplot(POTASS, main="Répartition de la quantité de potassium")
#Description de potass
summary(WEIGHT); boxplot(WEIGHT, main="Répartition des poids") #Description de potass
boxplot((cereal[,c(-1,-2,-3,-13,-14,-15)])) ####Affichage de la répartition des variables
boxplot(log(cereal[,c(-1,-2,-3,-13,-14,-15)])) ###Affichage du log de toutes les valeurs numériques importantes pour voir les disparités

#Commenter sur les recommandations de la FDA suivant les valeurs des nutriments
barplot(table(VITAMINS)) #Histogramme
table(VITAMINS) #Valeurs

#CAH sur les barres de céréales, on distingue 6 groupes
#centrage réduction des données
#pour éviter que variables à forte variance pèsent indûment sur les résultats
cah.cereal <- scale(cereal[,-1:-3],center=T,scale=T)
#matrice des distances entre individus
cah.mat <- dist(cah.cereal)
#CAH - critère de Ward
#method =  ward.D2  correspond au vrai critère de Ward
#utilisant le carré de la distance
cah.ward <- hclust(cah.mat,method="ward.D2")
#affichage dendrogramme
plot(cah.ward,xlab = "Alias of Names",sub = "CAH")
```