

# TD1

Nom des étudiants : SY Abiboulaye et LEGRAND Axel

## 1.1 Chargement des données

Après avoir téléchargé le jeu de données, nous effectuons les commandes suivantes pour ouvrir le fichier et avoir certaines caractéristiques de position pour chaque colonne du dataframe avec la commande summary()

```
> Data <- read.csv ("D:/home/coursera/datasets_910_1662_menu.csv")
> df <- as.data.frame(Data)
> summary(df)
```

	Category	Item	Serving.Size
Coffee & Tea	:95	1% Low Fat Milk Jug	: 1 16 fl oz cup: 45
Breakfast	:42	Apple Slices	: 1 12 fl oz cup: 38
Smoothies & Shakes	:28	Bacon Buffalo Ranch McChicken	: 1 22 fl oz cup: 20
Beverages	:27	Bacon Cheddar McChicken	: 1 20 fl oz cup: 16
Chicken & Fish	:27	Bacon Clubhouse Burger	: 1 21 fl oz cup: 7
Beef & Pork	:15	Bacon Clubhouse Crispy Chicken Sandwich	: 1 30 fl oz cup: 7
(Other)	:26	(Other)	:254 (Other) :127

Calories	Calories.from.Fat	Total.Fat	Total.Fat....Daily.Value.
Min. : 0.0	Min. : 0.0	Min. : 0.000	Min. : 0.00
1st Qu.: 210.0	1st Qu.: 20.0	1st Qu.: 2.375	1st Qu.: 3.75
Median : 340.0	Median : 100.0	Median : 11.000	Median : 17.00
Mean : 368.3	Mean : 127.1	Mean : 14.165	Mean : 21.82
3rd Qu.: 500.0	3rd Qu.: 200.0	3rd Qu.: 22.250	3rd Qu.: 35.00
Max. : 1880.0	Max. : 1060.0	Max. : 118.000	Max. : 182.00

Saturated.Fat	Saturated.Fat....Daily.value.	Trans.Fat	Cholesterol
Min. : 0.000	Min. : 0.00	Min. : 0.0000	Min. : 0.00
1st Qu.: 1.000	1st Qu.: 4.75	1st Qu.: 0.0000	1st Qu.: 5.00
Median : 5.000	Median : 24.00	Median : 0.0000	Median : 35.00
Mean : 6.008	Mean : 29.97	Mean : 0.2038	Mean : 54.94
3rd Qu.: 10.000	3rd Qu.: 48.00	3rd Qu.: 0.0000	3rd Qu.: 65.00
Max. : 20.000	Max. : 102.00	Max. : 2.5000	Max. : 575.00

Cholesterol....Daily.value.	Sodium	Sodium....Daily.value.	Carbohydrates
Min. : 0.00	Min. : 0.0	Min. : 0.00	Min. : 0.00
1st Qu.: 2.00	1st Qu.: 107.5	1st Qu.: 4.75	1st Qu.: 30.00
Median : 11.00	Median : 190.0	Median : 8.00	Median : 44.00
Mean : 18.39	Mean : 495.8	Mean : 20.68	Mean : 47.35
3rd Qu.: 21.25	3rd Qu.: 865.0	3rd Qu.: 36.25	3rd Qu.: 60.00
Max. : 192.00	Max. : 3600.0	Max. : 150.00	Max. : 141.00

Carbohydrates	Daily value	Dietary Fiber	Dietary Fiber	Daily value
---------------	-------------	---------------	---------------	-------------

### 1. Quelles conditions devez-vous respecter pour utiliser un test du coefficient de corrélation linéaire ?

Les conditions sont les suivantes :

- Les variables doivent être quantitatives continues
- Le vecteur composé des variables suit une loi gaussienne

### 2. Justifiez l'utilisation d'un test du coefficient de corrélation linéaire sur ces variables.

Nous utilisons un test de corrélation linéaire pour savoir si les variables sont linéairement dépendantes ou indépendantes entre elles

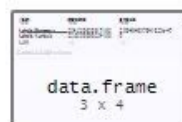
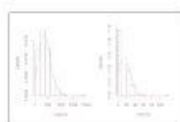
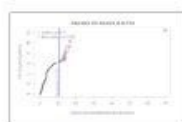
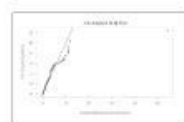
### 3. Comment pouvez-vous conclure sur la corrélation linéaire des deux variables Calories et Total.Fat ?

Vérifions si les conditions mentionnées à la question sont respectées

-Le vecteur composé des deux variables suit-elle une loi normale?

Nous utilisons le test de multinormalité "Mardia" pour effectuer le test. avec l'option, mvnTest = "mardia". Par défaut, le test de Henze-Zirkler est appliqué Et nous utilisons le test de Shapiro-Wilk dans l'option univariateTest="SW"

```
library(MVN)
result <- mvn(data = data_macdo[,c("Calories", "Total.Fat")],
  mvnTest = "mardia", univariateTest = "Sw",
  univariatePlot = "histogram", multivariatePlot = "qq",
  multivariateOutlierMethod = "adj", showOutliers = TRUE, showNewData = TRUE)
result$multivariateNormality
```



Test <fctr>	Statistic <fctr>	p value <fctr>	Result <fctr>
Mardia Skewness	224.229958917198	2.30464057341325e-47	NO
Mardia Kurtosis	22.0280366224199	0	NO
MVN	NA	NA	NO

3 rows

Après utilisation de la fonction mvn du package MVN, on se rend compte que les deux tests de multinormalité (Mardia Skewness et Mardia Kurtosis) sont significatifs au seuil  $\alpha = 0.05$ . Ce qui peut être vérifié avec la dernière ligne MVN qui contient la valeur NO. Donc on rejette l'hypothèse de binormalité.

Cependant on peut vérifier pourquoi le test est significatif en visualisant les outliers avec la commande result\$multivariateOutliers. On obtient le résultat suivant:

```
result$multivariateOutliers
```

	Observation <fctr>	Mahalanobis Distance <dbl>	Outlier <chr>
83	83	69.638	TRUE
33	33	16.556	TRUE
250	250	15.838	TRUE
247	247	14.853	TRUE
32	32	14.173	TRUE
82	82	13.683	TRUE
244	244	13.498	TRUE
252	252	13.498	TRUE
254	254	13.231	TRUE
35	35	12.802	TRUE

1-10 of 11 rows

Previous 1 2 Next

On voit qu'il y a 11 variables atypiques. L'observation 83 étant la plus atypique car sa distance est beaucoup trop grande par rapport aux autres.

on va voir à quoi correspond l'observation 83

```
## {r}
data_macdo[83,c("Item","Calories","Total.Fat")]
```

	Item <fctr>	Calories <int>	Total.Fat <dbl>
83	Chicken McNuggets (40 piece)	1880	118

1 row

L'observation correspond au 40 piece de chicken McNuggets qui est très riche en calories.

On suppose la binormalité du couple car les outliers ne sont que 11 dans l'observation totale

Appliquons la fonction cor

```
## {r}
cor.test(data_macdo$Calories,data_macdo$Total.Fat)
```

Pearson's product-moment correlation

```
data: data_macdo$Calories and data_macdo$Total.Fat
t = 34.048, df = 258, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8795250 0.9243604
sample estimates:
      cor
0.9044092
```

Soit  $H_0 : \rho(X,Y) = 0$

et  $H_1 : \rho(X,Y) \neq 0$   
avec  $\rho$  : coefficient de corrélation linéaire.

On voit que le test est significatif au seuil  $\alpha = 0.05$  car la p-value est inférieure à  $\alpha$ , on rejette donc l'hypothèse nulle et on accepte l'hypothèse  $H_1$  avec une intervalle de confiance comprise entre [0.88;0.92]. Donc les variables sont fortement corrélées entre elles et le coefficient de corrélation est égale à 0.90,

**4. Comment testeriez-vous l'indépendance de variables explicatives quantitatives deux à deux ? Si vous avez une idée, pouvez-vous présenter vos résultats sous forme de tableau pour les variables suivantes : Calories, Total.Fat, Cholesterol, Sodium, Sugars et Protein**

On peut utiliser le test de nullité du coefficient de corrélation linéaire avec la condition de normalité bivariée. La condition d'application du test est que la normalité bivariée de tous les couples doit être vérifiée.

Le cardinal de l'ensemble des couples possibles est égale à 15. Comme la plupart des couples ne suivent pas une loi normale bivariée, il faut faire un test de permutation.

```
set.seed(1133)
r_c_mdo <- perm.cor.mtest(data_macdo[,c("Calories", "Total.Fat", "Cholesterol", "Sodium", "Sugars", "Protein")], num.sim = 50000)
lapply(r_c_mdo, round, 4)
```

```
$p
  [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  0 0.0000 0.000  0 0.0000 0.000
[2,]  0 0.0000 0.000  0 0.0618 0.000
[3,]  0 0.0000 0.000  0 0.0280 0.000
[4,]  0 0.0000 0.000  0 0.0000 0.000
[5,]  0 0.0618 0.028  0 0.0000 0.004
[6,]  0 0.0000 0.000  0 0.0040 0.000

$cor
  [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 1.0000 0.9044 0.5964 0.7123 0.2596 0.7878
[2,] 0.9044 1.0000 0.6805 0.8462 -0.1154 0.8078
[3,] 0.5964 0.6805 1.0000 0.6244 -0.1355 0.5616
[4,] 0.7123 0.8462 0.6244 1.0000 -0.4265 0.8698
[5,] 0.2596 -0.1154 -0.1355 -0.4265 1.0000 -0.1799
[6,] 0.7878 0.8078 0.5616 0.8698 -0.1799 1.0000
```

Nous allons utiliser la fonction `rquery.format` qui calcule simultanément les coefficients de corrélation linéaire et les p-valeurs associées au test de nullité du coefficient de corrélation linéaire une fois que les hypothèses de normalité multivariée sont vérifiées.

Comme la normalité des multivariée n'est pas vérifiée dans plusieurs couples, nous n'allons afficher que le tableau des corrélations `$r`. On obtient le tableau ci-dessous



```
source("http://www.sthda.com/upload/rquery_cormat.r")
rquery.cormat(data_macdo[,c("Calories", "Total.Fat", "Cholesterol", "Sodium",
                             "Sugars", "Protein")])$r
```



The screenshot shows an RStudio window. On the left, a 'data.frame' with 6 rows and 6 columns is displayed. On the right, a correlation matrix is shown as a heatmap and a table. The table below represents the correlation matrix data.

	Sugars <fctr>	Cholesterol <fctr>	Calories <fctr>	Total.Fat <fctr>	Sodium <fctr>	Protein <fctr>
Sugars	1					
Cholesterol	-0.14	1				
Calories	0.26	0.6	1			
Total.Fat	-0.12	0.68	0.9	1		
Sodium	-0.43	0.62	0.71	0.85	1	
Protein	-0.18	0.56	0.79	0.81	0.87	1

A travers le tableau, on voit qu'il n'existe aucune couple de variables indépendants. Mais les couples de variables (Sugars, Cholesterol), (Calories Sugars), (Total.Fat, Sugars) et (Sugars, Protein) ont un coefficient de corrélation inférieur à 0,5, on peut en conclure qu'ils sont peu dépendants.

## 1.2 Corrélation linéaire entre deux variables

### 5. Déterminez deux groupes de variables entre elles qui présentent des corrélations linéaires supérieures en valeur absolue à 0,5.

Voici deux groupes de variables entre elles qui présentent des corrélations linéaires supérieures en valeur absolue à 0,5 :

a) "Protein", "Calories", "Total.Fat", "Sodium", "Cholesterol", "Saturated.Fat", "Total.Fat....Daily.Value.", "Cholesterol....Daily.Value." and "Saturated.Fat....Daily.Value.".

```
> list <- c("Protein", "Calories", "Total.Fat", "Sodium", "Cholesterol", "Saturated.Fat", "Total.Fat....Daily.value.", "cholesterol....Daily.value.", "Saturated.Fat....Daily.value.")
> round(cor(df[, list]), 2)
```

	Protein	Calories	Total.Fat	Sodium	Cholesterol	Saturated.Fat	Total.Fat....Daily.value.	cholesterol....Daily.value.	Saturated.Fat....Daily.value.
Protein	1.00	0.79	0.81	0.87	0.56	0.60	0.81	0.56	0.61
Calories	0.79	1.00	0.90	0.71	0.60	0.85	0.90	0.60	0.85
Total.Fat	0.81	0.90	1.00	0.85	0.68	0.85	1.00	0.68	0.85
Sodium	0.87	0.71	0.85	1.00	0.62	0.58	0.85	0.62	0.59
Cholesterol	0.56	0.60	0.68	0.62	1.00	0.63	0.68	1.00	0.63
Saturated.Fat	0.60	0.85	0.85	0.58	0.63	1.00	0.85	0.63	1.00
Total.Fat....Daily.value.	0.81	0.90	1.00	0.85	0.68	0.85	1.00	0.68	0.85
cholesterol....Daily.value.	0.56	0.60	0.68	0.62	1.00	0.63	0.68	1.00	0.63
Saturated.Fat....Daily.value.	0.61	0.85	0.85	0.59	0.63	1.00	0.85	0.63	1.00

b) "Sugars", "Calcium....Daily.Value." and "Carbohydrates"

```
> list2 <- c("Sugars", "Calcium....Daily.value.", "Carbohydrates")
> round(cor(df[, list2]), 2)
```

	Sugars	Calcium....Daily.value.	Carbohydrates
Sugars	1.00	0.60	0.76
Calcium....Daily.value.	0.60	1.00	0.59
Carbohydrates	0.76	0.59	1.00

### 6. Justifiez l'utilisation d'une ACP pour ce jeu de données.

L'analyse en composantes principales se justifie pour ce jeu de données car on cherche à établir des profils nutritionnels pour les différents menus, en cherchant des ressemblances ou des oppositions entre ces différents profils nutritionnels. Les objectifs de l'ACP sont :

- établir un bilan des ressemblances entre les individus (lignes du tableau)
- réaliser un bilan des corrélations entre les variables (colonnes du tableau)

-Mettre en liaison les deux études : trouver les variables caractéristiques d'un groupe d'individus donnés.

Dans ce jeu de données, l'utilisation des groupes de variables précédemment établis peuvent former nos composantes principales d'ACP, ceux-ci ayant des variables très corrélées entre elles.

**7. Expliquez les différences qu'il y a entre une ACP normée et une ACP non normée ?**

En Analyse en composantes principales non normées, on n'effectue pas de réduction(diviser chaque variable par son écart type) donc les variables n'auront pas la même variance alors qu'en analyse en composantes principales normées, les variables sont réduites donc leur variance sera égale à 1. L'ACP ne favorise alors plus une variable dont la variance est plus élevée qu'une autre.

**8. Quel type d'ACP utiliseriez-vous ici ? Justifiez votre réponse**

Nous allons utiliser une ACP normée pour équilibrer l'influence des variables dans le calcul des ressemblances afin de s'affranchir de l'unité de mesure car plusieurs variables n'ont pas la même unité de mesure ou la même ordre de grandeur. La variable « Protein » varie entre 0 et 87 tandis que la variable « Calcium » varie entre 0 et 3600.

### **1.3 Représentation en trois dimensions**

Représentation en 3D des trois variables Calories, Total.Fat, Cholesterol pour chaque individu

**9. À quoi sert l'option type="s" dans la fonction plot3d ?**

L'argument type dans plot3d est utilisé pour indiquer le type d'élément que trace la fonction plot3d . L'option type ="s" permet d'afficher les données en sphères alors que l'option type="p" affiche les données en points.

**10. Quelles différences voyez-vous entre ce graphique et le plot en 3D précédent ?**

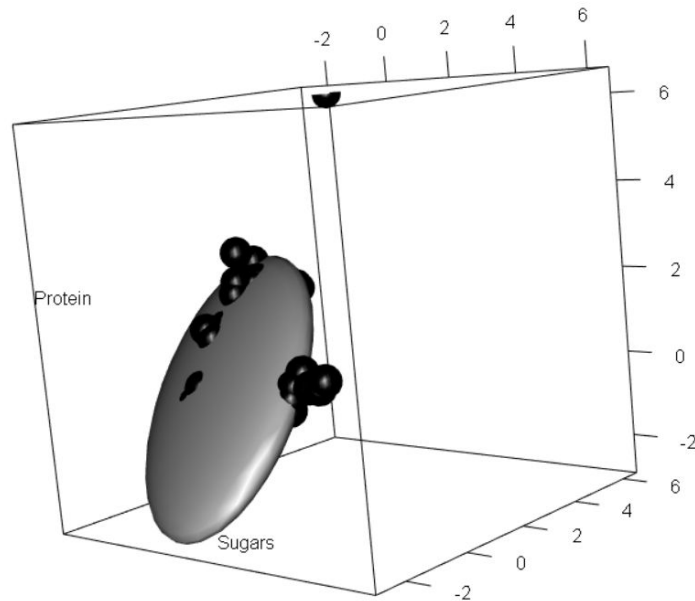
L'échelle est plus petite dans le cas normé (nouveau graphique) car les variables ont été centrées et réduites en divisant les données centrées par l'écart-type. Les échelles sont aussi les mêmes sur les 3 axes dans le cas normé (valeurs allant d'environ 0 à environ 6) alors qu'avant normalisation, les plages allaient de 0 à 1200 (axe Total.Fat), 0 à 2000 (axe Calories) et 0 à 600 (axe Cholestérol).

**11. Commentez la répartition des points dans l'ellipse**

Les points du graphe se trouvent majoritairement dans l'ellipse créée, à l'exception de quelques outliers et de certains points n'étant pas dans l'ellipse mais en étant très proches.

Cette ellipse contient environ 90% de l'ensemble des points.

**12. Affichez l'ellipse de corrélation linéaire dans la représentation en 3D pour les attributs Sodium, Sugars et Protein.**



Code R:

```
listQ12 <- c("Sugars", "Sodium", "Protein")
df.cr <- scale(data_macdo[, listQ12])
lims <- c(min(df.cr), max(df.cr))
plot3d(df.cr, type="s", xlim=lims, ylim=lims, zlim=lims)
df2 <- as.data.frame(df.cr)
plot3d(ellipse3d(cor(cbind(df2$Sugars, + df2$Sodium, df2$Protein))), col="grey",
add=TRUE)
```

**13. Expliquez les différences entre les ellipses obtenues dans les deux nuages.**

L'ellipse obtenue avec le premier nuage est une ellipse majoritairement construite selon un axe. En effet celle-ci est dirigée par l'axe  $x = y = z$ . C'est une ellipse d'un nuage créé avec des variables corrélées.

La seconde ellipse est quant à elle construite pour contenir la majorité des données, données formant deux axes perpendiculaires. En prenant l'axe du "Sodium" étant l'axe x, l'axe "Sugars" étant l'axe y et l'axe "Protein" tant l'axe z, cette seconde

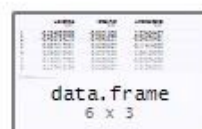
ellipse est alors dirigée par les axes  $x = z$  et  $y = 0$ . C'est une ellipse d'un nuage créé avec des variables non corrélées.  
Cela peut-être vérifié en construisant les tableaux de corrélations des deux groupes de variables.

## 1.4 Analyse en Composantes Principales (ACP)

### 14. Que contient le dataframe tab ?

le dataframe tab contient les données des 3 variables (Calories, Total.Fat, Cholesterol) centrées et réduites et a pour dimension 260 lignes et 3 colonnes

```
library(ade4)
list <- c("Calories", "Total.Fat", "Cholesterol")
acp <- dudi.pca(data_macdo[, list], center=TRUE, scale=TRUE, scannf = FALSE, nf = 3)
names(acp)
head(acp$tab)
```



	Calories <dbl>	Total.Fat <dbl>	Cholesterol <dbl>
1	-0.284683606	-0.0821929	2.35424457
2	-0.493184275	-0.4348357	-0.34376431
3	0.007217331	0.6230927	-0.11414653
4	0.340818401	0.9757356	2.64126679
5	0.132317732	0.6230927	-0.05674209
6	0.257418134	0.6230927	2.81348012

### 15. Comparez-le avec le tableau de données data\_macdo.cr. Expliquez la légère différence

```
data_macdo.cr <- scale(d_macdo[, list])
data_macdo.cr_df <- as.data.frame(data_macdo.cr)
head(data_macdo.cr_df)
```

	Calories <dbl>	Total.Fat <dbl>	Cholesterol <dbl>
1	-0.284135610	-0.08203469	2.34971281
2	-0.492234930	-0.43399870	-0.34310258
3	0.007203438	0.62189333	-0.11392681
4	0.340162350	0.97385733	2.63618253
5	0.132063030	0.62189333	-0.05663286
6	0.256922622	0.62189333	2.80806437

6 rows

data\_macdo.cr fournit les données centrées et réduites avec la fonction scale.



En faisant une comparaison avec les 2 tableaux de données `acp$tab` et `data_macdo.cr`, on voit que les données sont sensiblement égales.

Comme la fonction `scale` utilise la variance corrigée, donc les données sont réduites avec  $1/\sqrt{n-1}$  contrairement à la fonction `dudi.pca` qui utilise la variance empirique, c'est à dire les données sont réduites en  $1/\sqrt{n}$ , ce qui explique cette légère différence.

#### 16. Quelle manipulation devez-vous réaliser pour retrouver exactement le tableau utilisé dans `dudi.pca()` ?

Il suffit de multiplier le tableau de utilisé dans `scale()` par  $\sqrt{n/(n-1)}$  avec  $n=260$  car

$$\frac{dudi.pca()}{\sqrt{n}} = \frac{scale()}{\sqrt{n-1}}$$

```
head((data_macdo.cr_df[,list])*sqrt(260/259))
```

	Calories <dbl>	Total.Fat <dbl>	Cholesterol <dbl>
1	-0.284683606	-0.0821929	2.35424457
2	-0.493184275	-0.4348357	-0.34376431
3	0.007217331	0.6230927	-0.11414653
4	0.340818401	0.9757356	2.64126679
5	0.132317732	0.6230927	-0.05674209
6	0.257418134	0.6230927	2.81348012

6 rows

On obtient le même tableau utilisé dans `dudi.pca()`

### 1.5 Informations associées à une ACP

#### 17. Que vaut le pourcentage de l'inertie total avec 3 axes ?

Le pourcentage de l'inertie totale avec 3 axes est de 100%.

```
Decomposition of total inertia:
      inertia    cum  cum(%)
Ax1  2.46246    2.462   82.08
Ax2  0.44959    2.912   97.07
Ax3  0.08795    3.000  100.00
```

#### 18. Cherchez la signification du vecteur `rank`

Il donne le rang de la matrice diagonalisée. Dans notre cas, c'est le nombre de variables indépendantes

#### 19. Cherchez la signification du vecteur `nf`.

C'est le nombre de facteurs retenus dans l'analyse

#### 20. Cherchez la signification du vecteur `c1`.

Le vecteur `c1` donne les coordonnées des variables(colonnes) centrées et réduites

**21. Cherchez la signification du vecteur l1.**

l1 donne les coordonnées des individus (lignes) centrées et réduites

**22. Cherchez la signification du vecteur co**

co donne les coordonnées des variables (colonnes)..

**23. Cherchez la signification de l'objet call**

L'objet call conserve la façon dont les calculs ont été effectués lors de l'appel de la fonction dudi.pca()

**24. Cherchez la signification du vecteur cent**

cent donne les moyennes (cent pour centrage) des variables analysées

**25. Cherchez la signification du vecteur norm**

Ce vecteur donne les écarts-types (sur  $\sqrt{n}$ ) des variables analysée

**26. Donnez le nombre de facteurs retenus**

Question 26 : Donner le nombre de facteurs retenus

```
```{r}  
acp$nf|
```

```
[1] 3
```

On conclut que le nombre de facteurs retenus est 3

## 1.6 Analyse des variables

**27. Comment reconnaissez-vous sur la figure qu'un attribut est bien représenté ?**

Différents arguments permettent de reconnaître et de vérifier qu'un attribut est bien représenté, notamment avec le tableau "Coordonnees des attributs" :

	Comp 1	Comp 2	Comp 3
Calories	-0.93	-0.31	0.19
Total.Fat	-0.96	-0.18	-0.22
Cholesterol	-0.82	0.56	0.04

Ce tableau nous permet de voir quels variables doivent être les plus proches et selon quels axes. On voit donc grâce à la première colonne du tableau, que selon

le premier axe, en abscisse, la variable "Total.Fat" est la plus proche de la variable "Calories", ce qui se vérifie sur la figure. On voit aussi, avec la seconde colonne du tableau, que "Calories" est la variable la plus proche de "Total.Fat", ce qui se vérifie une nouvelle fois sur la figure.

On peut aussi vérifier que les attributs sont bien représentés grâce aux valeurs du tableau. Prenons "Total.Fat" qui a -0.96 selon l'axe 1 et -0.18 selon l'axe 2, valeurs que l'on peut vérifier sur la figure, avec l'attribut "Total.Fat" se trouvant dans la partie inférieure gauche de la figure, partie où les valeurs des axes 1 et 2 sont négatives. On peut faire de même avec les autres attributs.

On peut aussi voir si un attribut est bien représenté en regardant la longueur de sa représentation sur le cercle de corrélation. Plus cette dernière est proche de 1, plus la représentation est bonne et pleine d'informations.

**28. Quel est l'attribut le moins bien représenté dans le cercle de corrélation ? Justifiez votre réponse.**

L'attribut le moins bien représenté est "Total.Fat" car c'est l'attribut qui contient le plus d'information selon l'axe 3, qui n'est pas représenté ici. En effet, avec le tableau "Signed column relative contributions", 4.9% de "Total.Fat" contribue à l'axe 3, pour seulement 3.8% par "Calories" et 0.14% par "Cholesterol".

**29. A l'aide de la figure précédente (Figure 5), précisez l'attribut le plus corrélé positivement à Calorie ?**

L'attribut le plus positivement corrélé à "Calories" est l'attribut Total.Fat car c'est l'attribut qui en est le plus proche sur la figure et dont les valeurs des composantes sont les plus proches.

**30. Quels sont les attributs qui ont contribué à la construction de l'axe F1 ? Justifiez votre réponse.**

Les attributs ayant contribué à la construction de l'axe F1 sont les 3 attributs, qui ont contribué quasi équitablement, avec les pourcentages suivants : Calories à 35.06%, Total.Fat à 37.31% et Cholesterol à 27.63%. Cette information est fournie par le tableau intitulé "Column absolute contributions (%)".

**31. Donnez une signification à cet axe.**

Cet axe est un axe construit, quasi équitablement, par les 3 attributs et qui est donc l'axe principal de cet ACP. Cependant on peut voir que 2 variables sont principalement corrélées à cet axe, "Calories" et "Total.Fat". Cet axe nous donne donc de l'information sur une variable que l'on pourrait nommer "General.Fat" et qui serait le nom de cet axe.

**32. Quels sont les attributs qui ont contribué à la construction de l'axe F2 ? Justifiez votre réponse.**

Les attributs ayant contribué à la construction de l'axe F2 sont les 3 attributs, mais principalement l'attribut "Cholesterol" avec 70.748% et l'attribut "Calories" avec 22.006%. L'attribut "Total.Fat" a aussi contribué à construire F2, avec un pourcentage de 7.245.

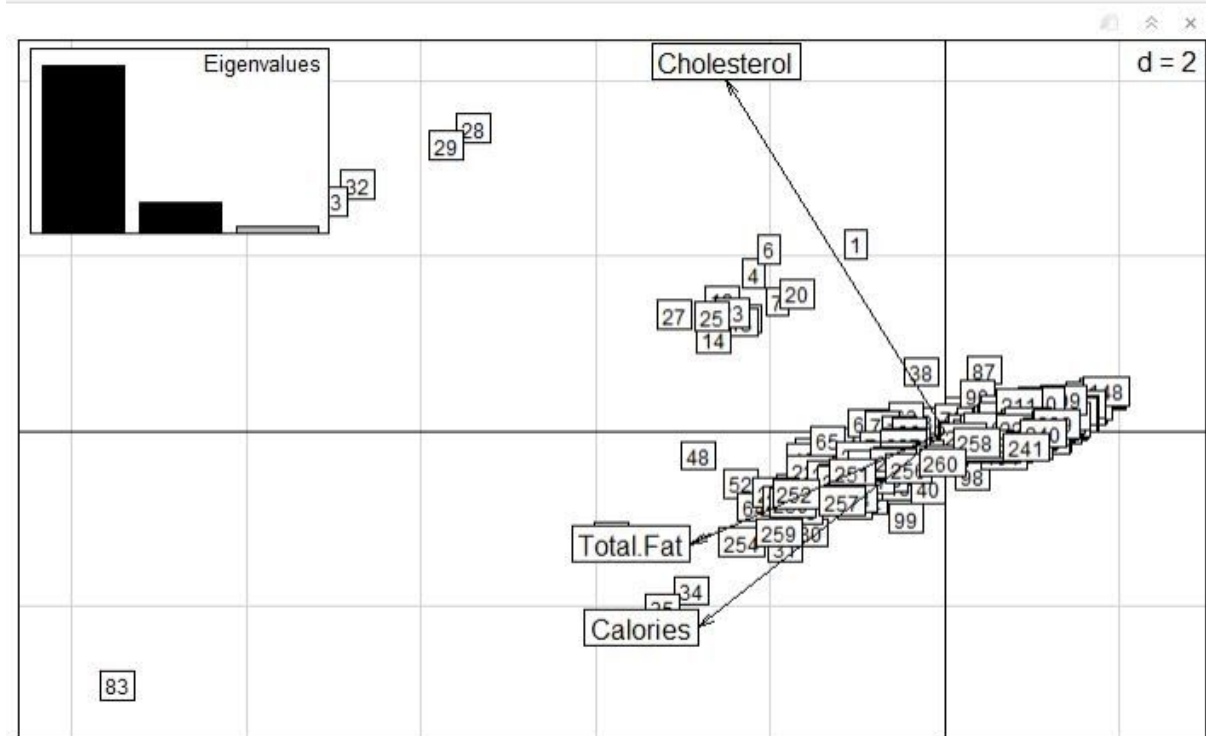
### 33. Donnez une signification à cet axe.

Cet axe est l'axe secondaire de notre ACP et est celui qui nous permet de faire la distinction entre des variables corrélées, plutôt corrélées et pas du tout corrélées. L'attribut majoritairement corrélé à cet axe est l'attribut "Cholesterol" et c'est la signification qu'à cet axe : la quantité de cholestérol dans l'aliment.

## 1.7 Conclusion

### 34. Utilisez la fonction scatter(acp).


```
#34. scatter() : renvoie une figure visuelle de synthèse de l'ACP
scatter(acp)
```



La fonction scatter donne une représentation simultanée, c'est à dire un biplot des individus et des variables. Elle affiche aussi le diagramme en barres des valeurs propres

### 35. Reprenez l'analyse à partir de la section 1.4 mais en incluant les variables Sodium, Sugars et Protein.

### 35.14 : Que contient le dataframe tab



The screenshot shows an R console window on the left and a data frame window on the right. The data frame window displays a 6x6 matrix of centered and scaled data.

	Calories <dbl>	Total.Fat <dbl>	Cholesterol <dbl>	Sodium <dbl>	Sugars <dbl>	Protein <dbl>
1	-0.284683606	-0.0821929	2.35424457	0.4414709	-0.9230901	0.32107066
2	-0.493184275	-0.4348357	-0.34376431	0.4761983	-0.9230901	0.40875802
3	0.007217331	0.6230927	-0.11414653	0.4935619	-0.9580251	0.05800856
4	0.340818401	0.9757356	2.64126679	0.6324712	-0.9580251	0.67182011
5	0.132317732	0.6230927	-0.05674209	0.6671985	-0.9580251	0.67182011
6	0.257418134	0.6230927	2.81348012	0.8061077	-0.9230901	1.11025694

6 rows

Figure : Dataframe tab contenant les 6 variables

Le dataframe tab contient les 6 variables centrées et réduites

### 35.15. Comparez-le avec le tableau de données data\_macdo.cr. Expliquez la légère différence.

```
data_macdo.cr2 <- scale(data_macdo[,list2])
data_macdo.cr_df2 <- as.data.frame(data_macdo.cr2)
head(data_macdo.cr_df2)
```



The screenshot shows an R console window on the left and a data frame window on the right. The data frame window displays a 6x6 matrix of centered and scaled data, which is very similar to the one in Figure 35.14.

	Calories <dbl>	Total.Fat <dbl>	Cholesterol <dbl>	Sodium <dbl>	Sugars <dbl>	Protein <dbl>
1	-0.284135610	-0.08203469	2.34971281	0.4406211	-0.9213133	0.3204526
2	-0.492234930	-0.43399870	-0.34310258	0.4752816	-0.9213133	0.4079712
3	0.007203438	0.62189333	-0.11392681	0.4926118	-0.9561810	0.0578969
4	0.340162350	0.97385733	2.63618253	0.6312537	-0.9561810	0.6705269
5	0.132063030	0.62189333	-0.05663286	0.6659142	-0.9561810	0.6705269
6	0.256922622	0.62189333	2.80806437	0.8045560	-0.9213133	1.1081198

6 rows

Figure : data\_macdo avec les 6 variables

On conclut de la même manière que la question 15, les données sont sensiblement égales, ce qui est expliqué par la différence de calcul de variance utilisée par la fonction `scale()` et la fonction `dudi.pca()`

### 35.16 Quelle manipulation devez-vous réaliser pour retrouver exactement le tableau utilisé dans `dudi.pca()` ?



35.16 Quelle manipulation devez-vous réaliser pour retrouver exactement le tableau utilisé dans `dudi.pca()` ?

```
##{r}
head((data_macdo.cr_df2[,list2]) * sqrt(260/259))
```

	Calories <dbl>	Total.Fat <dbl>	Cholesterol <dbl>	Sodium <dbl>	Sugars <dbl>	Protein <dbl>
1	-0.284683606	-0.0821929	2.35424457	0.4414709	-0.9230901	0.32107066
2	-0.493184275	-0.4348357	-0.34376431	0.4761983	-0.9230901	0.40875802
3	0.007217331	0.6230927	-0.11414653	0.4935619	-0.9580251	0.05800856
4	0.340818401	0.9757356	2.64126679	0.6324712	-0.9580251	0.67182011
5	0.132317732	0.6230927	-0.05674209	0.6671985	-0.9580251	0.67182011
6	0.257418134	0.6230927	2.81348012	0.8061077	-0.9230901	1.11025694

6 rows

Même démarche que la question 165

### 35.17 Que vaut le pourcentage de l'inertie total avec 3 axes ?

Pour les variables : Cholesterol, Total.Fat, Calories, Sodium, Sugars, Protein

Le pourcentage de l'inertie totale avec 3 axes est de 95,64%.

```
Decomposition of total inertia:
      inertia    cum cum(%)
Ax1  4.00005   4.000   66.67
Ax2  1.23452   5.235   87.24
Ax3  0.50390   5.738   95.64
Ax4  0.18437   5.923   98.71
Ax5  0.06727   5.990   99.84
Ax6  0.00989   6.000  100.00
```

### 35.26 Donnez le nombre de facteurs retenus.

Le nombre d'axes retenus est 6 en faisant `acp$nf`. Mais si nous voulions uniquement garder 80% de l'inertie totale, nous pourrions garder 2 axes car les deux premiers décrivent 87% de l'inertie totale. Cela nous permettrait aussi d'interpréter le problème en 2D.

**35.28 Quel est l'attribut le moins bien représenté dans le cercle de corrélation ? Justifiez votre réponse.**

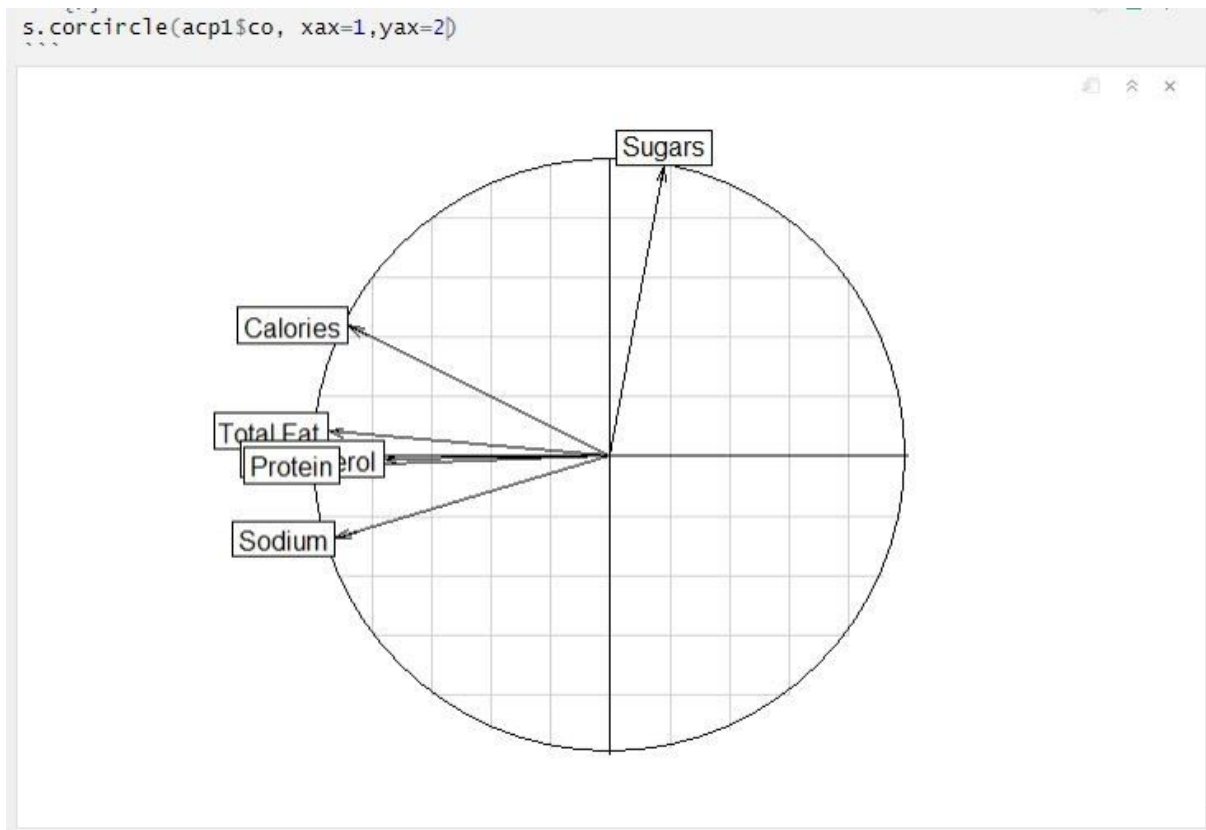


Figure: Cercle des corrélations des 6 variables sur les 2 premières axes

On commence par afficher les valeurs des  $\cos^2$  puis on additionne pour chaque variable son  $\cos^2$  sur chacun des deux axes utilisés pour créer le cercle des corrélations. Puis on les classe par ordre croissant pour identifier les valeurs les plus petites des  $\cos^2$  ( $\cos^2$  ou Qualité de représentation d'une variable). L'attribut le moins représenté est celui qui a le  $\cos^2$  le plus faible

```
Affichons les valeurs de cos2 par ordre croissant.
```{r}
round(sort(rowSums(get_pca_var(acp1)$cos2[,1:2])),2)
```
```

| Cholesterol | Protein | Total.Fat | Sodium | Calories | Sugars |
|-------------|---------|-----------|--------|----------|--------|
| 0.58        | 0.83    | 0.92      | 0.94   | 0.97     | 0.99   |

On voit bien aussi sur la figure qu'avec le calcul des  $\cos^2$  que l'attribut le moins représenté est le Cholesterol.

**35.29 À l'aide de la figure précédente (Figure de la question 35.28), précisez l'attribut le plus corrélé positivement à Calorie ?**

Total.Fat forme l'angle le plus petit avec le vecteur Calories. On peut en conclure que Total.Fat est l'attribut le plus corrélé à Calories comme à la question 29.

**35.30 Quels sont les attributs qui ont contribué à la construction de l'axe F1 ?**

**Justifiez votre réponse**

En utilisant l'élément \$col.abs qui donne le pourcentage des contributions , on obtient la sortie suivante:

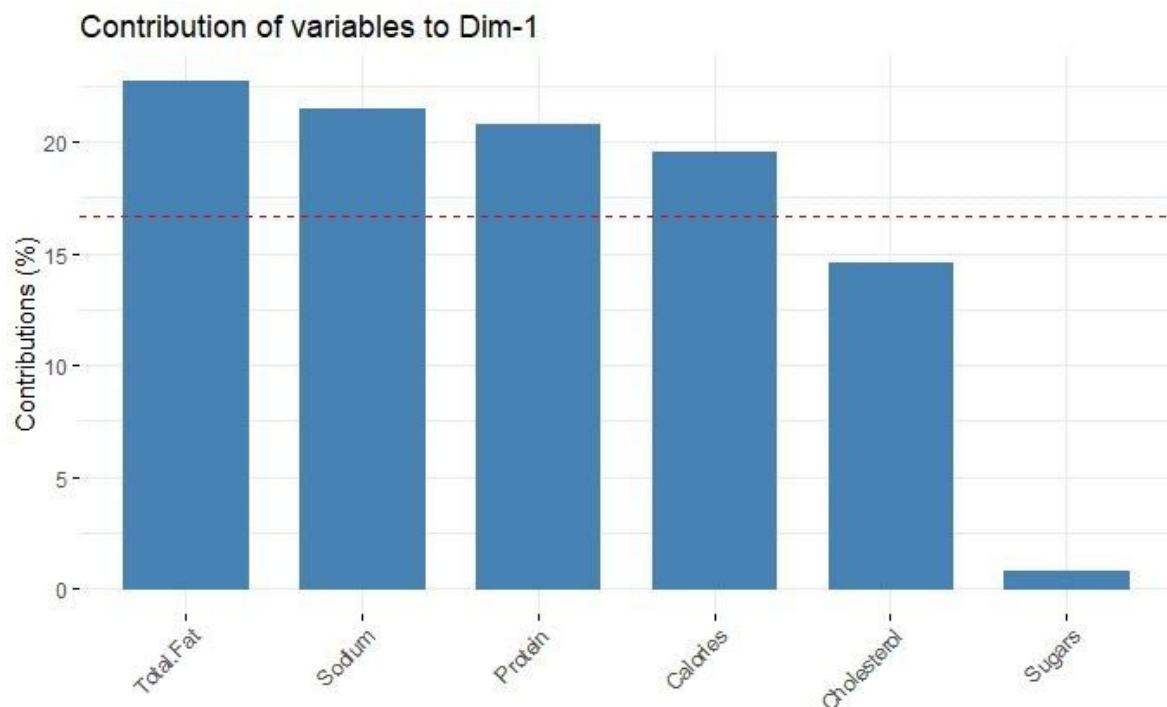
|             | Axis1<br><dbl> | Axis2<br><dbl> | Axis3<br><dbl> | Axis4<br><dbl> | Axis5<br><dbl> | Axis6<br><dbl> |
|-------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Calories    | 19.517         | 15.694         | 2.106          | 4.969          | 0.127          | 57.587         |
| Total.Fat   | 22.749         | 0.577          | 0.277          | 37.418         | 15.080         | 23.899         |
| Cholesterol | 14.594         | 0.008          | 80.968         | 4.400          | 0.003          | 0.026          |
| Sodium      | 21.530         | 6.195          | 3.722          | 0.008          | 64.049         | 4.496          |
| Sugars      | 0.823          | 77.450         | 0.018          | 3.234          | 5.362          | 13.113         |
| Protein     | 20.786         | 0.075          | 12.909         | 49.971         | 15.379         | 0.879          |

6 rows

Figure :

Total.Fat est l'attribut qui a le plus contribué à la construction de l'axe F1, suivi respectivement des attributs Sodium,Protein, Calories. L'attribut Cholesterol a une contribution en dessous de la moyenne des contributions et l'attribut Sugars a une contribution très faible sur la construction de l'axe F1

Ceci peut être mieux visualiser avec la fonction fviz\_contrib du package factoextra, nous obtenons un diagramme en barres qui donne le pourcentage des contributions des attributs.



### 35.31 Donner une signification à cet axe

Avec l'élément \$col.rel de la fonction inertia.dudi, nous obtenons les contributions relatives de la colonne signée.

|             | Axis1<br><dbl> | Axis2<br><dbl> | Axis3<br><dbl> | Axis4<br><dbl> | Axis5<br><dbl> | Axis6<br><dbl> |
|-------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Calories    | -78.070        | 19.374         | -1.061         | 0.916          | -0.009         | 0.570          |
| Total.Fat   | -90.998        | 0.712          | -0.140         | 6.899          | 1.014          | -0.236         |
| Cholesterol | -58.379        | -0.010         | 40.800         | -0.811         | 0.000          | 0.000          |
| Sodium      | -86.122        | -7.648         | -1.876         | -0.001         | -4.309         | -0.044         |
| Sugars      | 3.291          | 95.614         | 0.009          | -0.596         | -0.361         | -0.130         |
| Protein     | -83.146        | -0.093         | -6.505         | -9.213         | 1.035          | -0.009         |

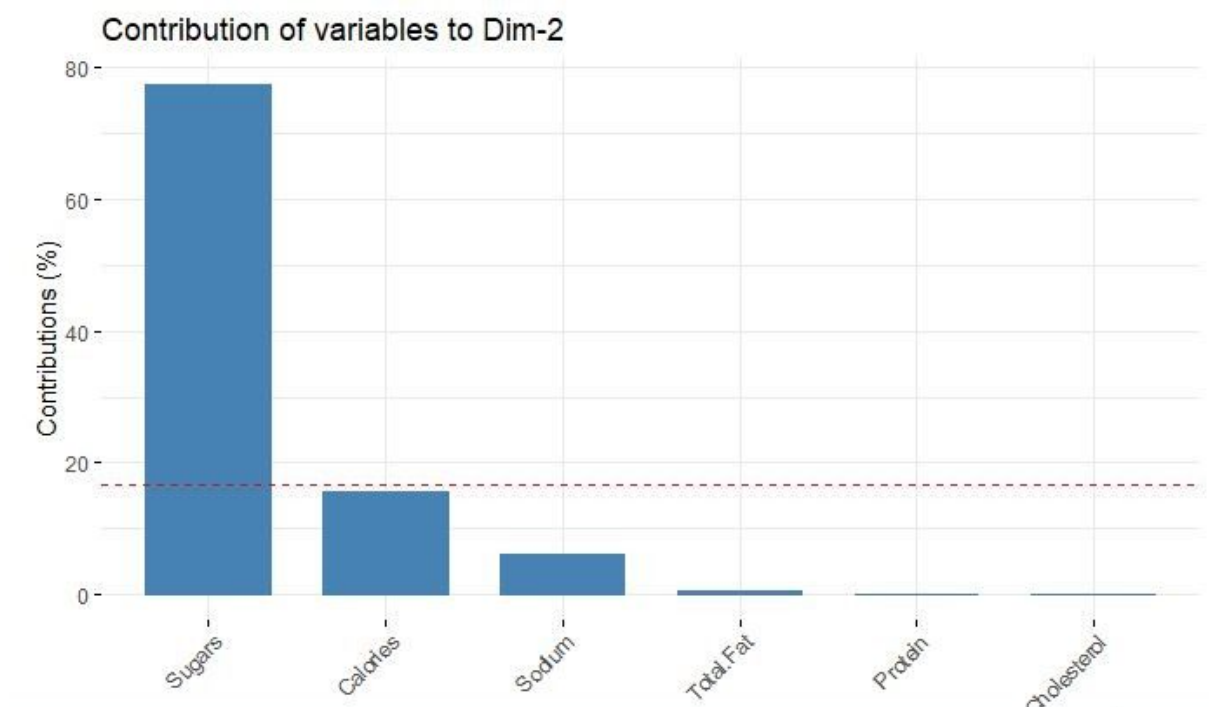
puis on s'intéresse aux coordonnées des variables sur les axes

|             | Comp1<br><dbl> | Comp2<br><dbl> | Comp3<br><dbl> | Comp4<br><dbl> | Comp5<br><dbl> | Comp6<br><dbl> |
|-------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Calories    | -0.884         | 0.440          | -0.103         | 0.096          | -0.009         | 0.075          |
| Total.Fat   | -0.954         | 0.084          | -0.037         | 0.263          | 0.101          | -0.049         |
| Cholesterol | -0.764         | -0.010         | 0.639          | -0.090         | 0.001          | 0.002          |
| Sodium      | -0.928         | -0.277         | -0.137         | -0.004         | -0.208         | -0.021         |
| Sugars      | 0.181          | 0.978          | 0.009          | -0.077         | -0.060         | -0.036         |
| Protein     | -0.912         | -0.030         | -0.255         | -0.304         | 0.102          | -0.009         |

6 rows

L'axe F1 est définie par la présence de toutes les variables ayant des coefficients négatifs, donc toutes sauf Sugars. On peut définir cet axe comme l'axe "Calories lipidiques d'aliments riches en protéines et sodium"

**35.32 Quels sont les attributs qui ont contribué à la construction de l'axe F2 ? Justifiez votre réponse**

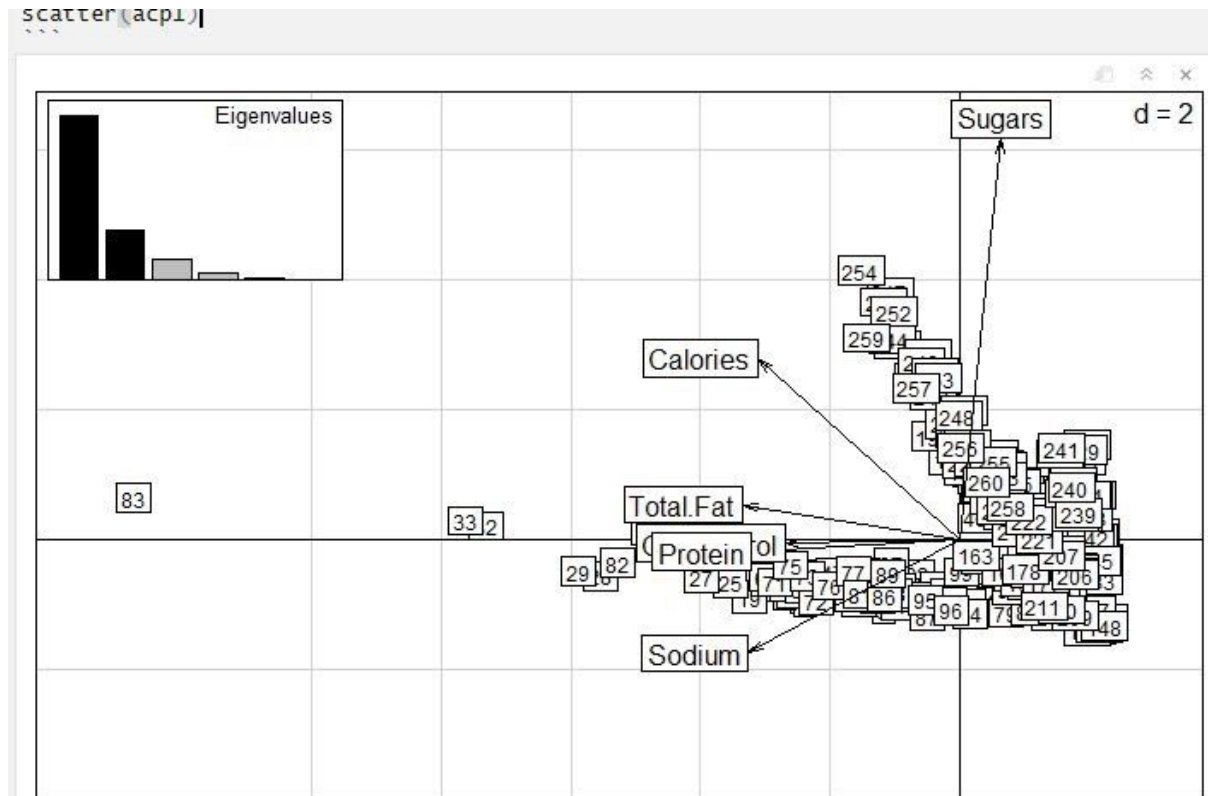


L'attribut Sugars est l'attribut prépondérant à la construction de l'axe F2. L'attribut Calories atteint de près la moyenne des contributions et l'attribut Sodium a une contribution faible. Les autres attributs ont une contribution très faible

**35.33 Donnez une signification à cet axe.**

on peut donner la signification de “Sucré” à cet axe par conclusion à la question précédente

### 35.34 Utiliser Scatter(acp1)



**36. Concluez sur le jeu de données. Iriez-vous prendre votre petit déjeuner chez MacDonald's ? Pourquoi ?**

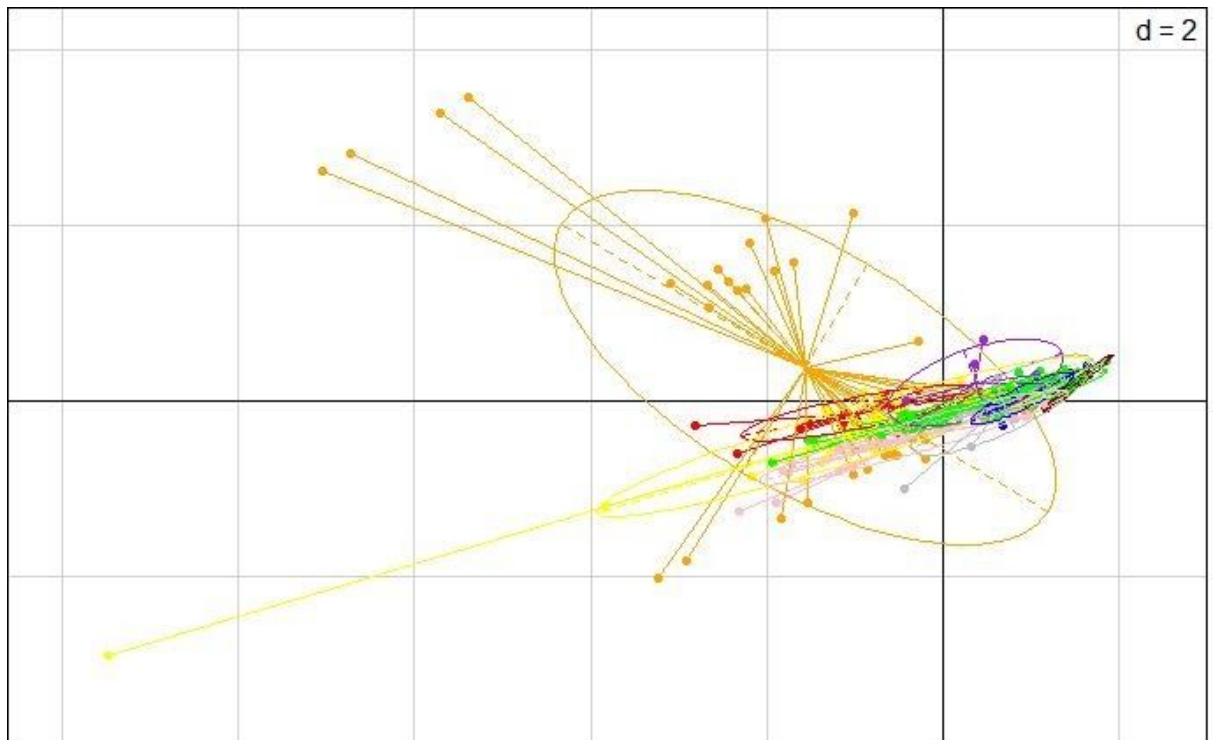
### ACP avec les variables Total.Fat, Calories, Cholestérol

La figure ci-dessous montre l'ACP en 2D représentée par classes, correspondant aux catégories de produits

| Classe         | Couleur     |
|----------------|-------------|
| Beef & Pork    | Rouge clair |
| Beverages      | Rouge foncé |
| Breakfast      | Orange      |
| Chicken & Fish | Jaune       |
| Coffee and Tea | Vert        |
| Desserts       | Bleu        |
| Salads         | Violet      |



|                    |      |
|--------------------|------|
| Smoothies & Shakes | Rose |
| Snacks & Sides     | Gris |



Interprétation:

Axe1 : General.Fat

On remarque que l'ellipse orange représentant les produits de petits-déjeuners(breakfast) est la plus étendue des ellipses sur l'axe du facteur 1 "General.Fat" avec la jaune, représentant elle la majorité des produits Poisson & Poulet. Son centre de gravité est aussi le plus à gauche de toutes les ellipses (celui des produits Poisson & Poulet est juste à sa droite). De plus, l'ellipse des boissons est quant à elle toute à droite, ce qui est en accord avec la réalité car les boissons sont trop riches en calories

On peut en déduire que :

- les points les plus à droites sur l'axe 1 correspondent aux produits faibles en calories grasses. Et, plus on se déplace vers la gauche, plus les produits sont riches en calories grasses.
- les petits-déjeuners vont du très riche au peu riches en calories grasses. même si la plupart sont riches en calories d'après l'ellipse
- les petits-déjeuners McDonald doivent sûrement être grandement composés de produits issus de poissons et poulets. Et ils fournissent beaucoup plus d'apports caloriques que les poissons&poulets, ce qui n'est pas bénéfique à la santé.

Axe2 : Cholesterol

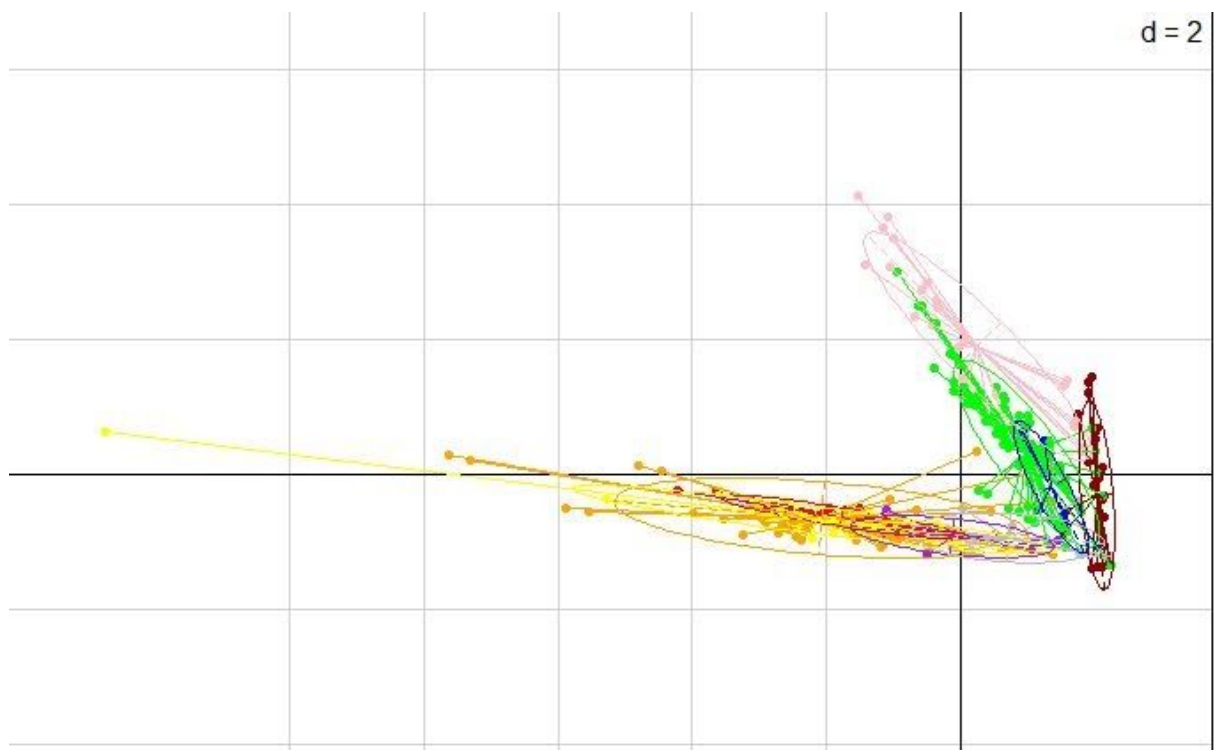
On remarque aussi que l'ellipse des produits de petits-déjeuners est de loin la plus étendue sur l'axe du facteur 2 "cholestérol". De plus, elle a les points les plus élevés sur l'axe 2 mais aussi les points les plus bas.

On en déduit que :

- On trouve de tout type de produits de petit-déjeuner par rapport au cholestérol, allant du moins au plus riche. Dans leur majorité, les petits-déjeuners McDonald sont plus riches en cholestérol que les produits des autres catégories mais on peut les éviter si on choisit bien ce qu'on commande.
- Une grande partie des produits de petits-déjeuners sont bien plus riches en cholestérol que les autres produits d'autres catégories.

Conclusion 1 : Un taux de cholestérol élevé peut provoquer des maladies cardiovasculaires comme le montre l'étude de Framingham, donc si je veux faire attention au taux de cholestérol dans mon sang et à la quantité de calories grasses que j'ingère, je n'irai pas prendre mon petit-déjeuner à McDonald. Mais c'est possible en étudiant et en sélectionnant précautionneusement les produits mais cela réduira grandement mon choix.

#### ACP avec les variables Total.Fat, Calories, Cholestérol, Sodium, Sugars, Protein



#### Interprétation

Axe F1: Calories lipidiques riches en protéines et sodium

On remarque que l'ellipse orange représentant les produits de petits-déjeuners est la plus étendue des ellipses sur l'axe du facteur 1. Son centre de gravité est l'un des

plus à gauche de toutes les ellipses et se confond avec les centres des ellipses rouge clair (Beef & Pork) et jaune (Chicken & Fish). Par contre l'ellipse des boissons est quant à elle toute à droite.

On peut en déduire que :

- les points les plus à droites sur l'axe 1 correspondent aux produits faibles en le facteur étudié. Et, plus on se déplace vers la gauche, plus les produits sont riches en le facteur étudié.
- les petits-déjeuners vont du très riche au peu riche en le facteur étudié. La tendance tend le plus vers des produits riches.
- les petits-déjeuners McDonald doivent surement être grandement composés de produits issus de poissons, poulets, boeufs et porcs

#### Axe F2: Sucre

On remarque aussi que l'ellipse des produits de petits-déjeuners fait partie de celles qui ont les étendues les plus faibles sur l'axe du facteur 2 "Sucre" contrairement aux ellipses rose (Smoothies & Shakes), verte (Coffee & Tea) et rouge foncée (Beverages) dont la majorité des points sont très hauts sur l'axe 2

On en déduit que :

- Dans leur majorité, les petits-déjeuners McDonald sont peu sucrés par rapport aux boissons et desserts.
- Les petits-déjeuners McDonald fournissent en moyenne les mêmes apports en sucre que les poissons, poulets, boeufs et porcs de McDonald. Ce qui paraît normal.

#### Conclusion:

Je peux prendre mon petit déjeuner au Macdo car le petit déjeuner est faible en sucre, ce qui n'est pas dangereux à ma santé. sauf si j'ai un problème avec les aliments riches en sodium, graisses et protéines.