République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي والبحث العلمي

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

المدرسة الوطنية العليا للإعلام الالي
Ecole nationale Supérieure d'Informatique

Rapport TP2 ANAD

2 -ème année Cycle Supérieur (2CS)

Option : Systèmes Intelligents et Données (SID)

Thème : AFCM/AFC

Réalisé par :

AIN GUERAD Manel

MAMMA Salima

Promotion : 2021 – 2022

Tables des matières

I. Introduction :

En analyse de données, il existe une panoplie de méthodes permettant le traitement et la visualisation de données à forte dimensionnalité en la réduisant tout en veillant à minimiser la perte d'information. Parmi lesquelles on cite l'ACP qui est dédiée aux variables continues, et ses analogues l'AFC et l'AFCM qui quant à elles manipulent des données catégorielles, l'AFC traite uniquement deux variables et l'AFCM est son extension sur plusieurs variables.

Ces techniques, nous permettent d'extraire des associations pertinentes entre les modalités des variables catégorielles, et d'identifier des groupes d'individus qui ont émis des réponses similaires au formulaire.

Pour bien comprendre l'utilité et le fonctionnement de l'AFCM et l'AFC, on les applique à travers un cas pratique dont nous discuterons en détail tout au long de ce travail.

II. Présentation du dataset choisi :

Pour notre étude, nous avons choisi le dataset tiré d'une enquête mené par starbuck, une chaîne de café multinationale, sur sa clientèle en leur posant une série de questions afin de comprendre les caractéristiques des clients récurrents de Starbucks et comment ces derniers se démarquent-ils des clients rares.
L'objectif de notre étude serait ainsi de répondre aux questions suivantes :
1- Qui sont les clients récurrents de Starbucks? Et qui sont les clients non fidèles?
2- Qu'est ce qui différencie les clients récurrents de ceux qui ne le sont pas ?
La compréhension de sa clientèle est d'une grande utilité pour Starbuck étant donné que ça va l'aider à mieux cibler sa stratégie de fidélisation des clients et renforcer également son plan Marketing

❖ **Caractéristiques clé du dataset :**

- Le nombre d'individus interrogés est de 113
- Le nombre de questions du formulaire est de 33, qu'on peut diviser en plusieurs catégories :
  1- Des questions relatives au clients (sexe, statut, salaire, age, sa position géographique par rapport à starbuck).
  2- Des questions relatives à sa consommation chez starbuck : nombre de visite, budget moyen dépensé chez starbuck, temps passé , possession d'une carte de crédit ou pas, achats et méthode de paiement
  3- Des questions relatives à son expérience avec starbuck : Evaluation service, le wifi, des produits, du prix

III. Réalisation :

Présentation des différents packages utilisés :

```
library("ade4")
library('dplyr')
library('FactoMineR')
library('factoextra')
```

1- Chargement du dataset sur R :

```
data <- read.csv("Data.csv", header = TRUE, sep = ",")#lecture des données
data
View(data)
```

2- Visualisation des propriétés du dataset :

```
> summary (data)
 gender age    status income visitNo method timeSpend location membershipCard itemPurchaseCoffee itempurchaseCold
 0:54   0:10   0:37   0:65   0: 2    0:44   0:64      0:25     0:60           1:113              1:113
 1:59   1:79   1:16   1:23   1: 9    1:20   1:34      1:32     1:53
        2:17   2:58   2:17   2:26    2:48   2:12      2:56
        3: 7   3: 2   3: 3   3:76    5: 1   3: 1
                      4: 5                  4: 2
 itemPurchasePastries itemPurchaseJuices itemPurchaseSandwiches itemPurchaseOthers spendPurchase productRate
 1:113                1:113              1:113                  1:113               0: 5          1: 1
                                                                                    1:56          2: 8
                                                                                    2:45          3:33
                                                                                    3: 7          4:48
                                                                                                  5:23
 priceRate promoRate ambianceRate wifiRate serviceRate chooseRate promoMethodApp promoMethodSoc promoMethodEmail
 1:12      1: 4      1: 1         1: 6     2: 4        1: 3       1:113          1:113          1:113
 2:25      2: 6      2: 5         2:13     3:36        2:15
 3:44      3:26      3:31         3:47     4:50        3:34
 4:23      4:41      4:50         4:37     5:23        4:40
 5: 9      5:36      5:26         5:10                 5:21
 promoMethodDeal promoMethodFriend promoMethodDisplay promoMethodBillboard promoMethodOthers loyal
 1:113           1:113             1:113              1:113                 0:  1              0:90
                                                                            1:112              1:23
```

- Notre dataset ne comporte pas de valeurs nulles ou manquantes, tous les individus ont répondu à toutes les questions posées.
- Nous avons constaté que certaines variables quantitatives disposent d'une seule modalité, comme par exemple ("itemPurchaseCoffe", "itempurchasecold"). Il ya aussi la variable "promoMethodOthers" qui est très mal distribuée, il ya un seul individu ayant la modalité 0. Ces variables ne nous serviront donc pas dans notre étude puisqu'elle n'apportent aucune information pertinente, nous avons donc décidé de les exclure de l'étude

Nous avons également exclu l'identifiant qui n'a aucune valeur à ajouter à notre étude

```
# enlever l'id
data = select(data , -1)

#selectionner que les variables à au moins deux modalités
data.new = select(data , -c(10 ,11,12,13,14,15,24,25,26,27,28,29,30,31))
```

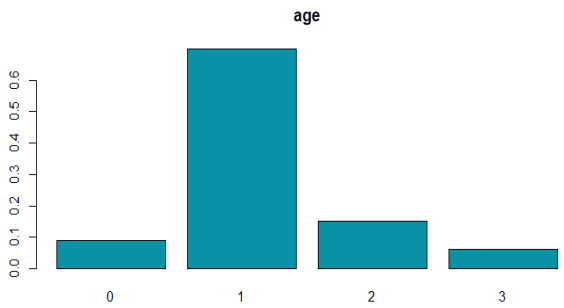
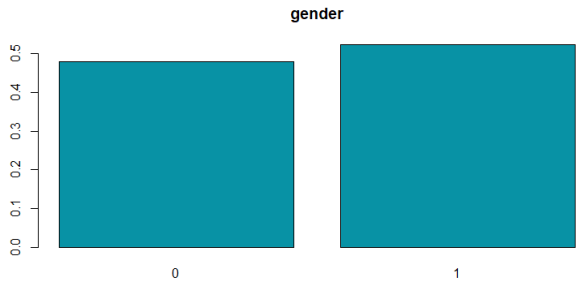
```
> summary(data.new)
 gender age    status income visitNo method timeSpend location membershipCard spendPurchase productRate priceRate promoRate ambianceRate wifiRate
 0:54   0:10   0:37   0:65   0: 2    0:44   0:64      0:25     0:60           0: 5          1: 1        1:12      1: 4      1: 1         1: 6
 1:59   1:79   1:16   1:23   1: 9    1:20   1:34      1:32     1:53           1:56          2: 8        2:25      2: 6      2: 5         2:13
        2:17   2:58   2:17   2:26    2:48   2:12      2:56                    2:45          3:33        3:44      3:26      3:31         3:47
        3: 7   3: 2   3: 3   3:76    5: 1   3: 1                              3: 7          4:48        4:23      4:41      4:50         4:37
                      4: 5                  4: 2                                          5:23        5: 9      5:36      5:26         5:10
 serviceRate chooseRate loyal
 2: 4        1: 3       0:90
 3:36        2:15       1:23
 4:50        3:34
 5:23        4:40
             5:21
```

- Etant donné l'objectif de notre étude, nous avons sélectionné les variables qui apportent le plus d'information pertinente, d'une part sur le profil des clients de par l'age, le sexe, le salaire et le statut social et d'une autre part sur la fréquence de sa visite aux cafés starbuks par la variable "visitNo"

```
#selectionner les variables pour afcm
data.new =select(data.new , c('gender', 'age', 'status', 'income' ,'visitNo'))
```

3- Visualisation des propriétés de chaque variable :

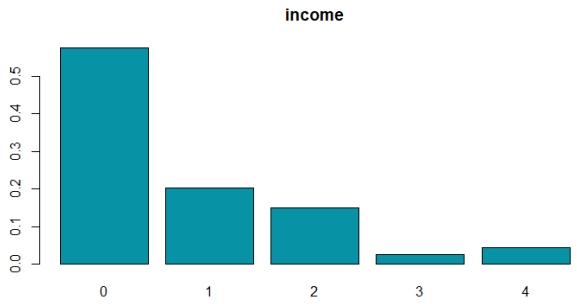
Nous allons à présent visualiser quelques caractéristiques des variables choisies

<table>
<tr>
<td>Age :
<pre>
         0   1   2   3
0.08849558 0.69911504 0.15044248 0.06194690
</pre>
0 : moins de 20 ans
1 : entre 20 et 29 ans
2 : entre 30 et 39 ans
3 : plus de 40 ans</td>
<td> 
</td>
</tr>
<tr>
<td>Gender :
<pre>
         0   1
0.4778761 0.5221239
</pre>
0 : Homme
1 : Femme</td>
<td> 
</td>
</tr>
<tr>
<td>Statut :
<pre>
         0   1   2   3
0.32743363 0.14159292 0.51327434 0.01769912
</pre>
0 : Etudiant
1 : Travailleur Indépendant
2: Employé
3 : Mére au foyer
On remarque que la modalité 3 est représentée par un seul individu uniquement on a donc décidé de l'exclure car elle n'apporte pas d'information pertinente</td>
<td> 
</td>
</tr>
</table>

income :

```
        0         1         2         3         4
0.57522124 0.20353982 0.15044248 0.02654867 0.04424779
```

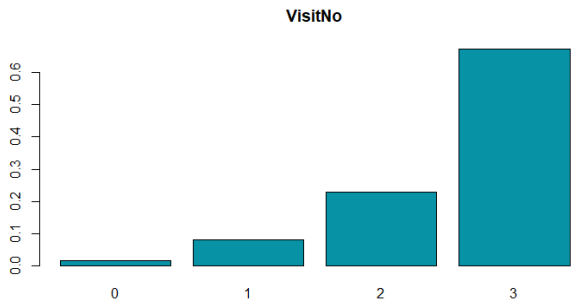
0 : <25.000 $
1 : entre 25.000 et 50.000
2 : entre 50.000 et 100.000
3 : entre 100.000 et 150.000
4 : >150.000
On a constaté que la variable n'est pas très bien distribué et que les modalités 3 et 4 sont très peu cités on a décidé donc de les jumeler pour avoir une modalité regroupant les salaires >100.000



visitNo

```
        0         1         2         3
0.01769912 0.07964602 0.23008850 0.67256637
```

0: Quotidiennement
1: Hebdomadaire
2: Mensuel
3: Rare



4 - Transformation du tableau de données en tableau disjonctif :

- Avant de construire le tableau disjonctif, nous avons renommé les modalités des variables : (Dans notre cas jamais est égale à des fréquence de visites rares)

```
  gender       age   status        income visitNo
1  Femme 20<age<29 étudiant salaire<25000  Jamais
2  Femme 20<age<29 étudiant salaire<25000  Jamais
3  Homme 20<age<29  employé salaire<25000 Mensuel
4  Femme 20<age<29 étudiant salaire<25000  Jamais
5  Homme 20<age<29 étudiant salaire<25000 Mensuel
6  Femme 20<age<29 étudiant salaire<25000  Jamais
```

```
#tableau disjonctif
data.disj = acm.disjonctif(data.new)
View(data.disj)
```

```
  gender.Femme gender.Homme age.<20 age.>40 ans age.20<age<29 age.30<age<39 status.employé
1            1            0      0           0             1             0              0
2            1            0      0           0             1             0              0
3            0            1      0           0             1             0              1
4            1            0      0           0             1             0              0
5            0            1      0           0             1             0              0
6            1            0      0           0             1             0              0
  status.étudiant status.indépendant income. 100000 < salaire income. 25000 < salaire < 50000
1               1                  0                        0                                0
2               1                  0                        0                                0
3               0                  0                        0                                0
4               1                  0                        0                                0
5               1                  0                        0                                0
6               1                  0                        0                                0
  income. 50000 < salaire < 100000 income.salaire<25000 visitNo.Hebdomadaire visitNo.Jamais
1                                0                    1                    0               1
2                                0                    1                    0               1
3                                0                    1                    0               0
4                                0                    1                    0               1
5                                0                    1                    0               0
6                                0                    1                    0               1
  visitNo.Mensuel visitNo.Quotidienne
1               0                   0
2               0                   0
3               1                   0
4               0                   0
5               1                   0
6               0                   0
```

5 - L'ACM :

- Avant de faire l'ACM, il est important de transformer les variables en facteur :

```
#transformer les valeurs des variables de chaines de caractère à facteur
i=0
while(i < ncol(data.new)){
  i=i+1
  data.new[,i] = as.factor(data.new[,i])
}
```

- Faire l'AFCM avec la fonction MCA du package FactoMineR:

```
#Faire l'afcm
mca = MCA(data.new)
```

6- Etude des valeurs propres :

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10
Variance               0.453   0.345   0.258   0.233   0.208   0.190   0.167   0.136   0.131   0.119
% of var.             18.887  14.391  10.749   9.718   8.672   7.915   6.949   5.668   5.454   4.949
Cumulative % of var.  18.887  33.278  44.027  53.746  62.418  70.333  77.282  82.949  88.404  93.353
                      Dim.11  Dim.12
Variance               0.091   0.068
% of var.              3.801   2.845
Cumulative % of var.  97.155 100.000
```

Pour savoir le nombre d'axes à retenir et pour pouvoir apprécier la qualité de représentation d'un axe, **Benzeckri** a proposé un taux d'inertie corrigé, Ce taux d'inertie est calculé comme suit :

1-On garde que les valeurs propres supérieurs au seuil (1/P) = 1/5 = 0.2

2-On corrige les inerties on utilisant la formule :

$$\widetilde{\lambda_k} = \left[\left(\frac{p}{p-1}\right)\left(\lambda_k - \frac{1}{P}\right)\right]^2 = \left[\left(\frac{5}{5-1}\right)\left(\lambda_k - \frac{1}{5}\right)\right]^2$$

```
#corriger les inerties
vp = mca$eig
vp = vp[,1]
vp = vp [vp>1/5]
vpa = 5/4 * (vp - 1/5)
vpa = vpa^2
I = vpa / sum(vpa) * 100
```

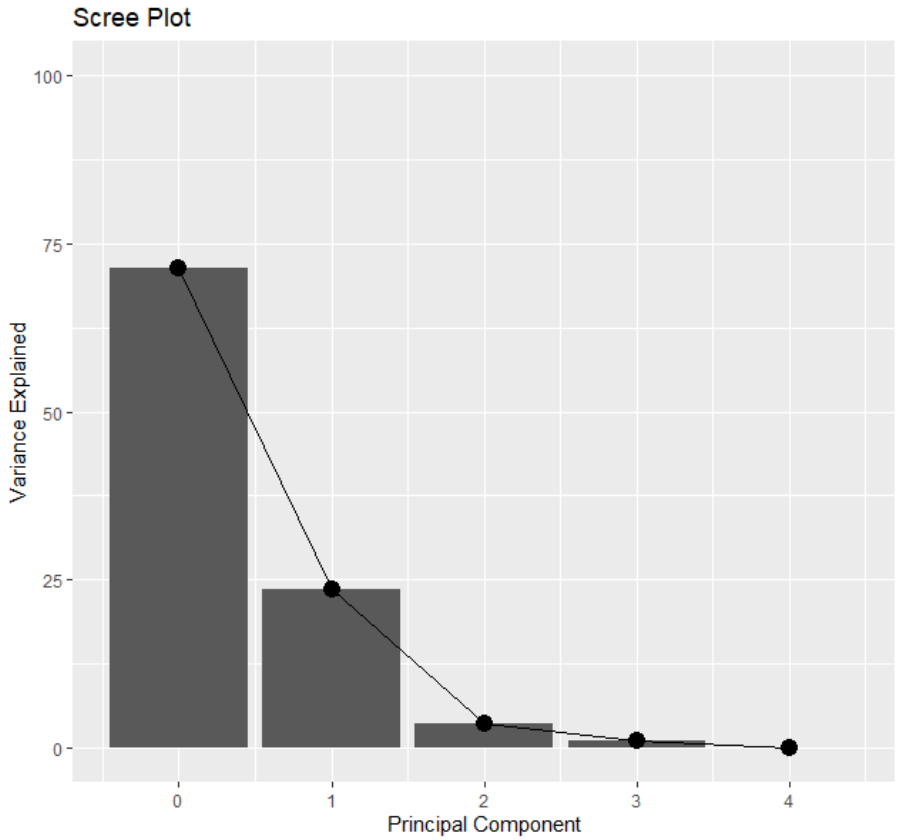
3-On obtient les inerties corrigées :

```
      dim 1       dim 2       dim 3       dim 4       dim 5
71.41972430 23.53332303  3.74342012  1.22999874  0.07353382
```

Méthode 1 : On voit qu'avec les 2 premiers axes on obtient un taux d'inertie très grand égal à 94.95%.

4-Utilisation de la méthode du coude :

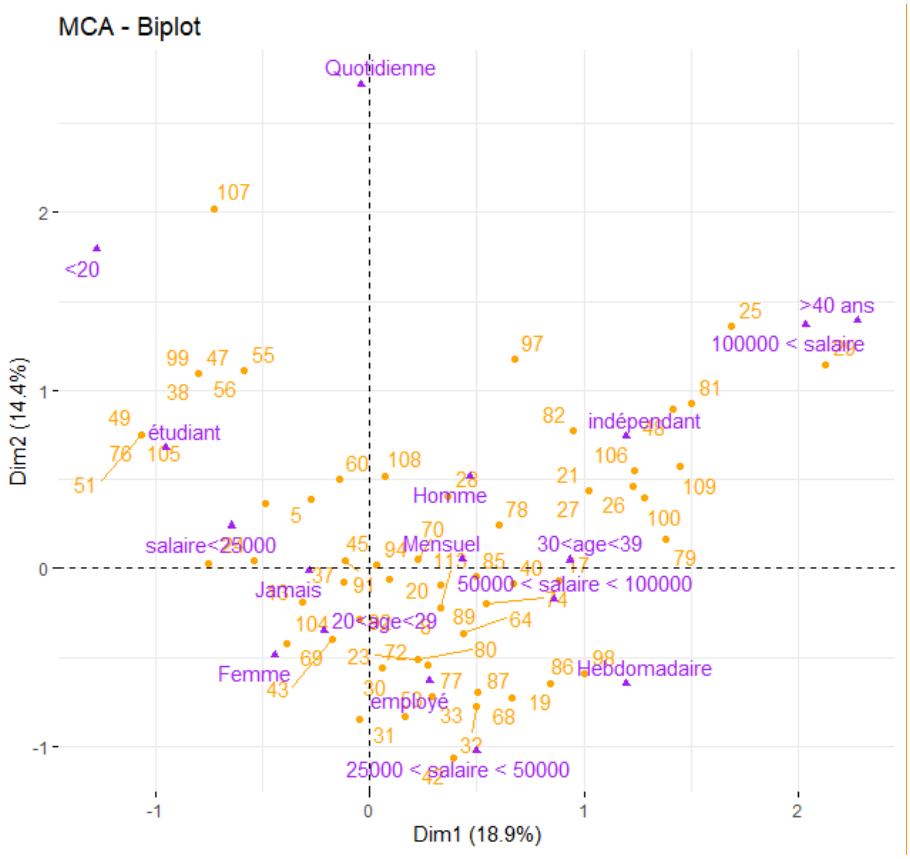
```
qplot(c(1:11), I) +
  geom_col()+
  geom_line() +
  geom_point(size=4)+
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 50)
```



Méthode 2 : On observe dans le graphe une brusque décroissance des valeurs propres corrigées à partir de la valeur propre 2.

Résultat : à partir des deux méthodes on décide de prendre 2 dimensions. Donc la qualité globale de la représentation est égale à 94.95%.

7- Représenter le biplot individus-variables :



8- Etudier le tableau des contributions et donnez une signification aux axes :

```
$coord
                                Dim 1       Dim 2       Dim 3       Dim 4       Dim 5
Femme                     -0.44312795 -0.48749983 -0.20948670  0.11115693  0.13703356
Homme                      0.46774617  0.51458315  0.22112485 -0.11733231 -0.14464654
<20                       -1.27799904  1.79063110  0.56584531  0.06467980  0.40934989
>40 ans                    2.27958867  1.39405141 -2.01136084  0.33602579 -0.25989392
20<age<29                 -0.21544289 -0.34698422 -0.15185001  0.23616727 -0.18471517
30<age<39                  0.93570611  0.04671470  1.07376546 -1.24023529  0.69844989
employé                    0.27887338 -0.63403118  0.20347701 -0.19584437 -0.14894437
étudiant                  -0.95476402  0.67438762  0.03005667 -0.14772504  0.26653386
indépendant                1.19697579  0.73884163 -0.80711020  1.05155000 -0.07643621
 100000 < salaire          2.03723289  1.36772918 -1.05725750 -1.31762352  0.40337393
 25000 < salaire < 50000   0.49844282 -1.02278353 -0.47901748  0.39559425  0.14214165
 50000 < salaire < 100000  0.85997730 -0.17297739  1.41426633  0.73658145 -0.01892756
salaire<25000             -0.64877632  0.23984103 -0.04926249 -0.16160911 -0.09677201
Hebdomadaire               1.19764344 -0.64690682  0.60369178  0.49963019  2.11954993
Jamais                    -0.28621424 -0.01237867 -0.31359318 -0.08448632  0.21387652
Mensuel                    0.43079292  0.05251617  0.52420495 -0.29825758 -1.41425494
Quotidienne               -0.04127312  2.71882866  2.49056924  4.64812103  0.11984236
```

```
$contrib
                              Dim 1        Dim 2       Dim 3       Dim 4       Dim 5
Femme                    4.449107741  7.066770722  1.74701939  0.54406918  0.92663157
Homme                    4.696280393  7.459369096  1.84407602  0.57429525  0.97811110
<20                      6.492333687 16.726668679  2.23617293  0.03231793  1.45066687
>40 ans                 12.393786494  6.082834980 16.95278972  0.52336260  0.35085027
20<age<29                1.439125659  4.899043163  1.25612887  3.36078126  2.30397675
30<age<39                5.916537587  0.019353222 13.68918905 20.20055531  7.17955912
employé                  1.793006662 12.163160160  1.67713727  1.71852710  1.11392136
étudiant                13.407061264  8.778455516  0.02334498  0.62375809  2.27554199
indépendant              9.112352110  4.556379693  7.27940016 13.66740464  0.08092758
 100000 < salaire       13.198091179  7.807058233  6.24541758 10.72948505  1.12689636
 25000 < salaire < 50000  2.271424478 12.551409293  3.68587407  2.78057103  0.40229891
 50000 < salaire < 100000 4.703636298  0.249744461 22.35076166  6.70606497  0.00496235
salaire<25000           10.708033899  1.920543467  0.10847324  1.29127235  0.51886909
Hebdomadaire             5.131417743  1.964821821  2.29077787  1.73558570 35.00323301
Jamais                   2.442212819  0.005995237  5.15115256  0.41356167  2.97006425
Mensuel                  1.844237719  0.035968578  4.79791093  1.71802396 43.28862203
Quotidienne              0.001354267  7.712423679  8.66437372 33.38036391  0.02486732
```

Une modalité contribue à la construction d'un axe si sa contribution absolue est supérieure à son poids*100.
Pour ce faire on a besoin des poids de chaque modalité calculé précédemment , on recalcule les poids des variables status et income car on a fait des modifications au niveau de ces variables :

- **Poids status :**

```
> nrow(data.new[data.new$status == 0,])/nrow(data.new)
[1] 0.3333333
> nrow(data.new[data.new$status == 1,])/nrow(data.new)
[1] 0.1441441
> nrow(data.new[data.new$status == 2,])/nrow(data.new)
[1] 0.5225225
```

- **Poids income :**

```
> nrow(data.new[data.new$income == 0,])/nrow(data.new)
[1] 0.5765766
> nrow(data.new[data.new$income == 1,])/nrow(data.new)
[1] 0.2072072
> nrow(data.new[data.new$income == 2,])/nrow(data.new)
[1] 0.1441441
> nrow(data.new[data.new$income == 3,])/nrow(data.new)
[1] 0.07207207
```

❖ *Axe 1 :*

+	-
>40 ans – salaire>100 000	

- L'axe 1 est un axe à effet "taille", il qualifie les clients de Starbucks qui sont payés plus de 100 000 qu'ils ont plus de 40 ans.

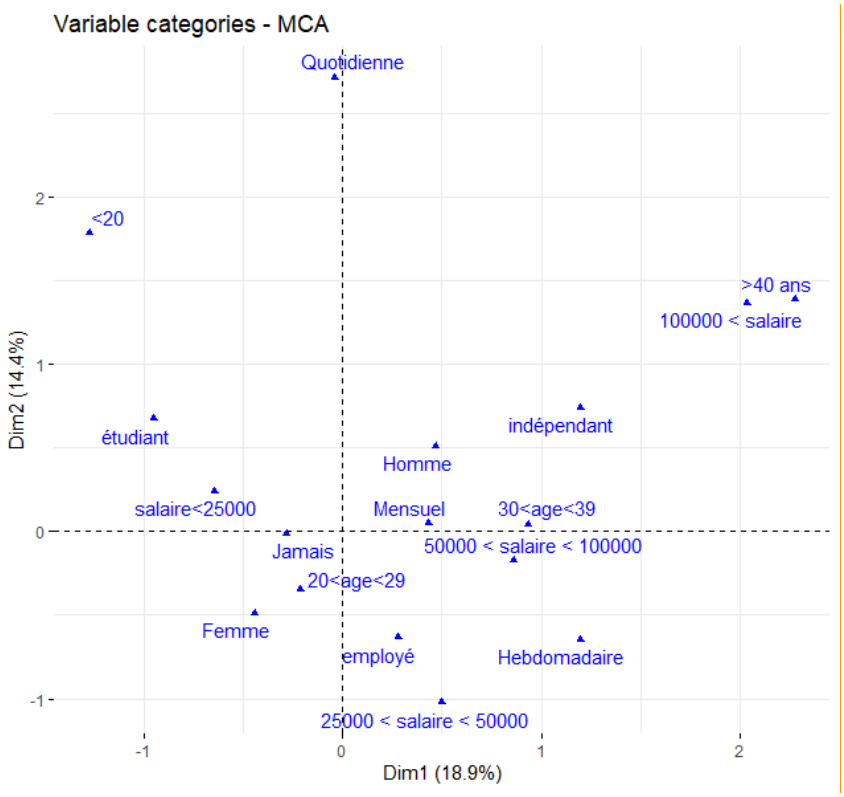
- ❖ *Axe 2 :*

+	-
Quotidienne – <20	

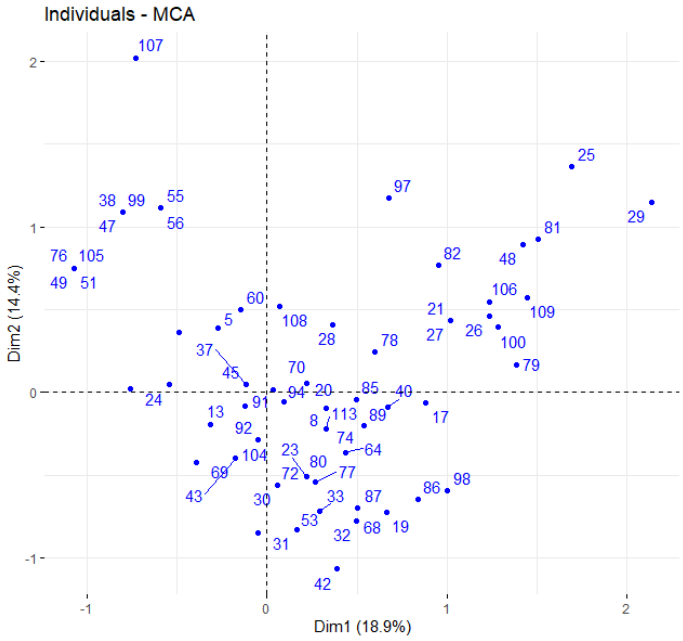
-L'axe 2 est un axe à effet "taille", il qualifie les clients de "Starbucks" qui ont moins de 20 ans et qu'ils viennent quotidiennement.

9 - Visualisations possibles :

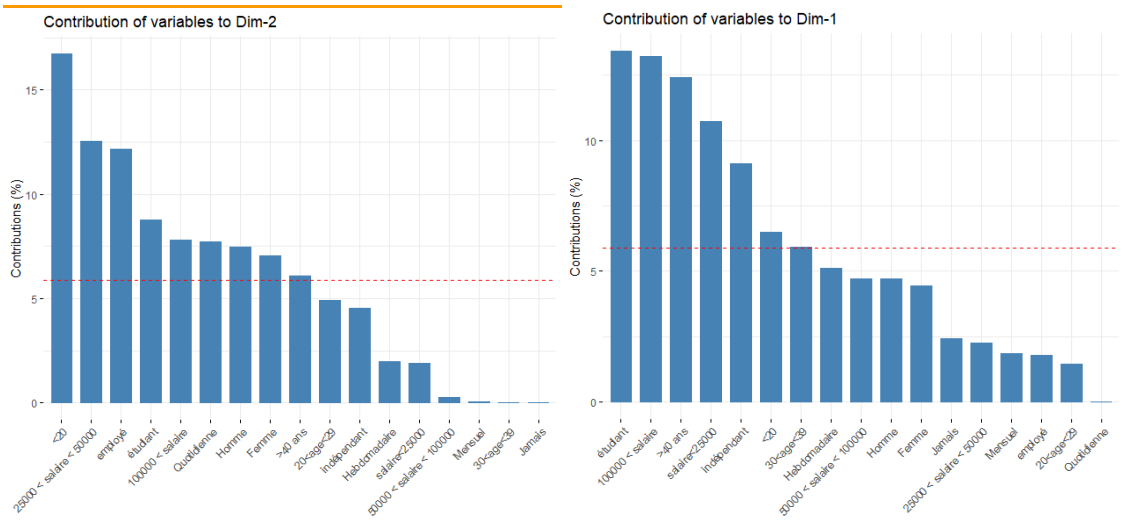
- Visualisation des variables :



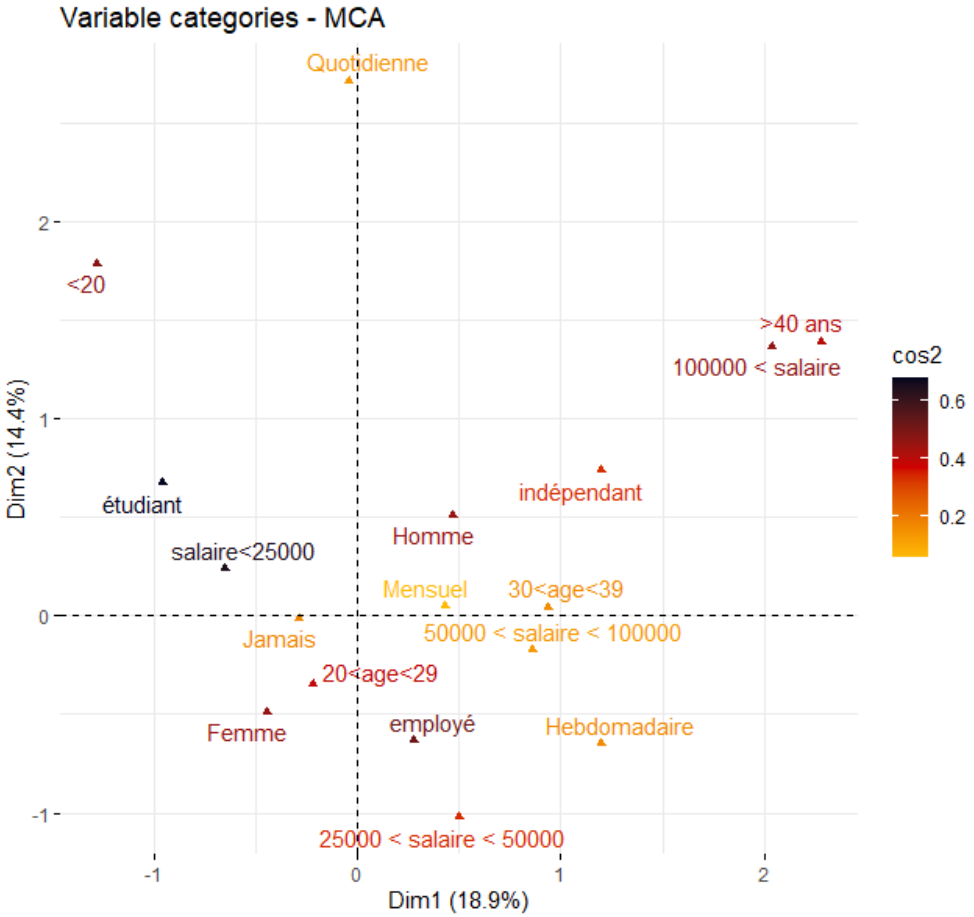
- Visualisation des individus :



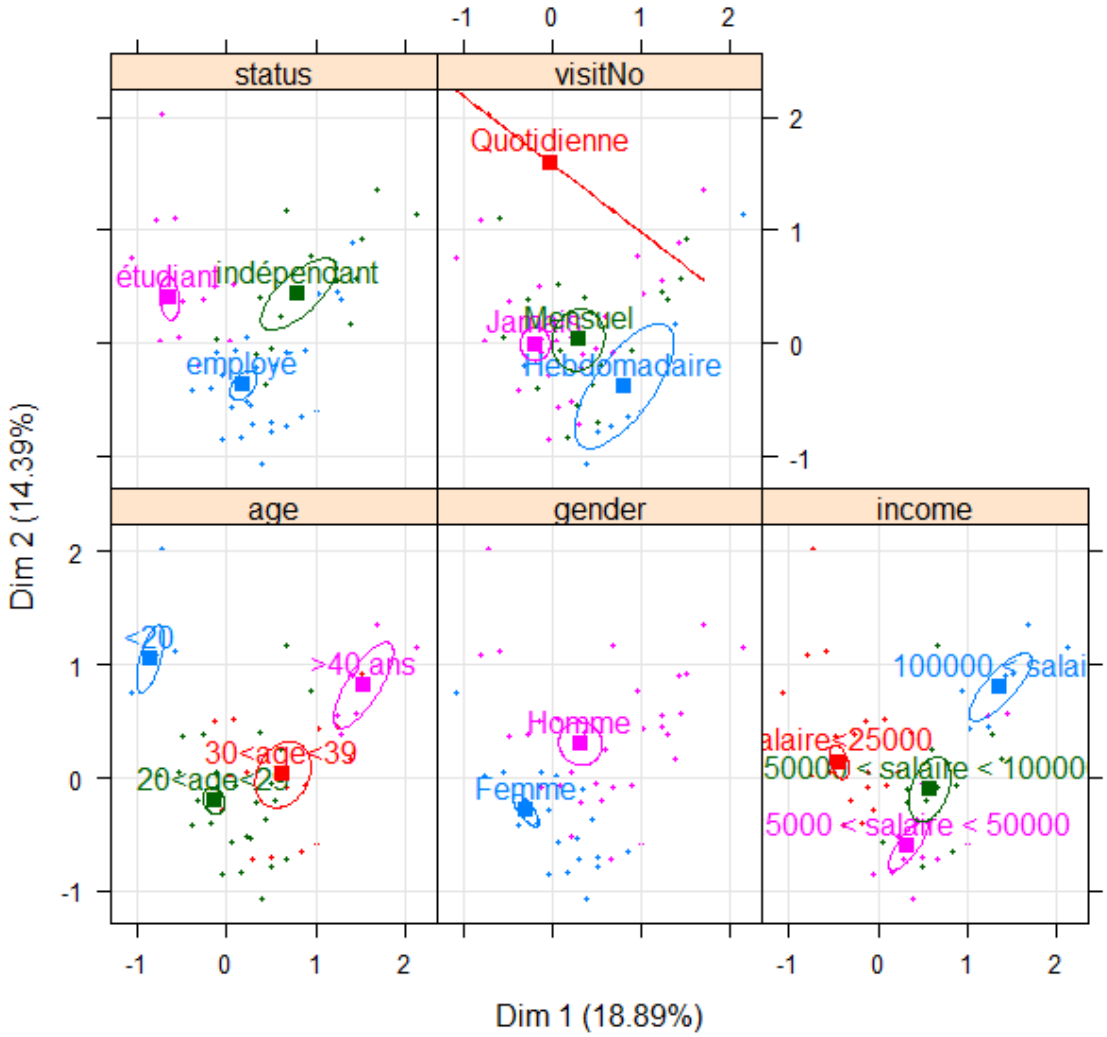
  - Visualisation des contributions absolues des variables sur les deux axes :



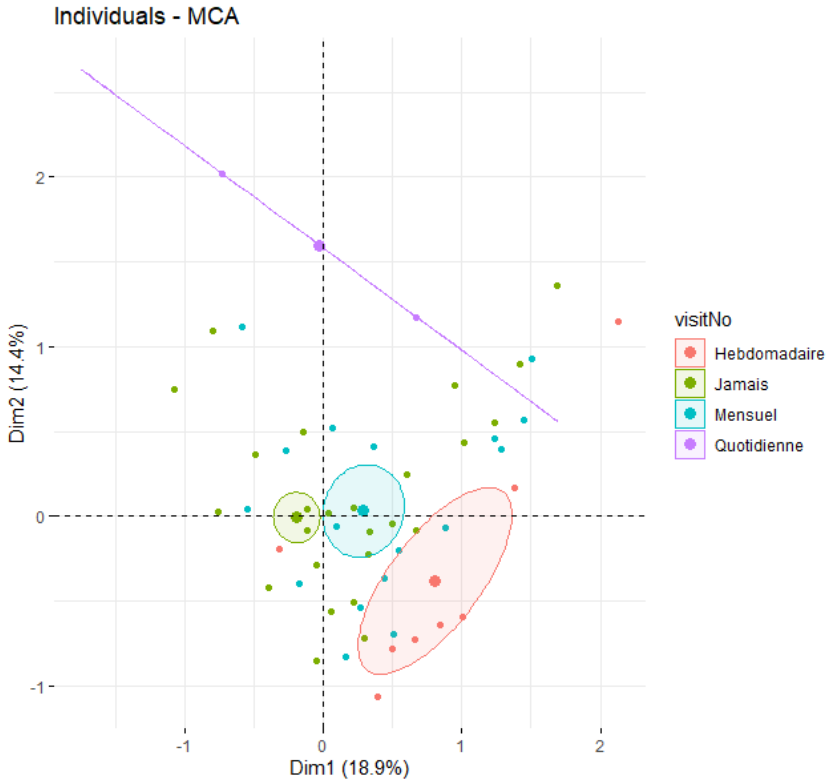
- Visualisation des contributions relatives des variables sur les deux axes pour percevoir celles qui sont les mieux représentées :



- Visualisation des variables en traçant des ellipses de confiance autour des modalités. L'objectif est de voir si les modalités d'une variable qualitative sont significativement différentes les unes des autres.



- Visualisation des modalités de la variables visitNo qui nous intéresse plus particulièrement dans notre cas , afin de voir où se situe les clients de “starbucks” aux visites régulières et ou se situe ceux qui sont moins fidèles afin de mieux cibler sa stratégie de commercialisation



10 - Ressortir les associations entre modalités :

A partir des visualisation précédentes et plus particulièrement le biplot variables individus et la visualisation des variables, on prélève quelques associations de modalités pertinentes qui nous serviront dans notre étude :

- Salaire compris entre 50.000 et 100.000 , âge compris entre 30 et 39, sexe masculin et visite mensuelle à Starbuck
- Age compris entre 20 et 29 , salaire <25.000, sexe féminin et visites rare chez starbucks (représentée par la modalité “jamais”)
- Âge supérieur à 40 ans ,salaire >100.000 et statut d’indépendant
- Visite hebdomadaire à starbuck, salaire compris entre 25000 et 50.000 et statut employé
- Statut d’étudiant et âge inférieur à 20 ans et salaire <25000

11- Interpréter les résultats :

- Les clients qui ont des visite hebdomadaire chez Starbuck sont majoritairement des employés dont le salaire est compris entre 25.000 et 50.000
- On constate que la majorité des clients à visite récurrentes de Starbucks sont des Hommes et que les femmes viennent rarement chez Starbucks
- Les personnes ayant un salaire <25.000 et sont âgés de 20 à 29 ans ont des visites rares chez Starbucks
- Les clients qui ont des visites mensuelles chez starbuck sont sont les personnes âgés de 30 à 39 ans avec un salaire compris entre 50.000 et 100.000 et sont généralement des hommes
- La clients âgés de moins de 20 ans sont des étudiants et ont un salaire <25.000
- Les clients qui ont le statut d'indépendant, sont majoritairement âgés de plus de 40 ans et ont un salaire >100.000
- Les employés âgés entre 20 et 29 ans et dont le salaire est entre 25000 et 50000 représentent une niche très importante des clients de Starbucks

12- Quelles sont les questions les mieux représentées par l'AFCM.

Pour savoir les questions les mieux représentées il faut calculer les contributions relatives :

```
$cos2
                                 Dim 1        Dim 2        Dim 3        Dim 4        Dim 5
Femme                        2.072714e-01 0.2508592011 0.0463227168 0.0130422990 1.982143e-02
Homme                        2.072714e-01 0.2508592011 0.0463227168 0.0130422990 1.982143e-02
<20                          1.617110e-01 0.3174613614 0.0317010809 0.0004142056 1.659082e-02
>40 ans                      2.969443e-01 0.1110502471 0.2311755683 0.0064521903 3.859706e-03
20<age<29                    1.097097e-01 0.2845772115 0.0545017336 0.1318317730 8.064655e-02
30<age<39                    1.583434e-01 0.0003946646 0.2085162594 0.2781821376 8.822498e-02
employé                      8.510719e-02 0.4399196381 0.0453088248 0.0419734146 2.427730e-02
étudiant                     4.557872e-01 0.2273993330 0.0004517018 0.0109113442 3.552015e-02
indépendant                  2.413054e-01 0.0919388561 0.1097139994 0.1862328255 9.839990e-04
 100000 < salaire            3.223548e-01 0.1452957756 0.0868189071 0.1348451838 1.263771e-02
 25000 < salaire < 50000     6.493455e-02 0.2734088784 0.0599719098 0.0409020529 5.280656e-03
 50000 < salaire < 100000    1.245576e-01 0.0050393562 0.3368672413 0.0913772190 6.033725e-05
salaire<25000                5.731550e-01 0.0783301713 0.0033045689 0.0355642601 1.275210e-02
Hebdomadaire                 1.265603e-01 0.0369254498 0.0321568030 0.0220262050 3.963963e-01
Jamais                       1.706637e-01 0.0003192322 0.2048764164 0.0148707053 9.529827e-02
Mensuel                      5.394841e-02 0.0008017290 0.0798810542 0.0258597636 5.814294e-01
Quotidienne                  3.125634e-05 0.1356335644 0.1138153233 0.3964225533 2.635264e-04
```

-Les questions les mieux représentées par l'AFCM sont celles qui ont les plus grandes contributions relatives (proches de 1).
-Les contributions relatives de chaque question par rapport à l'axe 1 et 2 :

cos2(gender) = 0.916
cos2(age) = 0.7267084(axe1) + 0.7134834846(axe2) = 1.44
cos2(income) = 1.085(axe1) + 0.5020741815(axe2) = 1.587
cos2(status) = 0.78219979(axe1) + 0.7592575671(axe2) = 1.541457357
cos2(visitNo) = 0.35(axe1) + 0.1737(axe2) = 0.5237

Donc les questions les mieux représentées par l'AFCM sont : age , income et status.

13- L'AFC :

```
The chi square of independence between the two variables is equal to 5.661689 (p-value =  0.2258782 ).
```

D'après le teste de KHI2 qui atteste s'il y a dépendance entre les 2 variables "VisitNo" et "status", on constate qu'il ya une dépendance entre ces 2 variables c-a-d entre les lignes et les colonnes (5.66 est beaucoup plus grande que 0.22), ainsi on peut utiliser ces 2 variables pour faire notre AFC.

Tableau de contingence :

```
             visitNo
status        Hebdomadaire Jamais Mensuel Quotidienne
  employé                6     36      16           0
  étudiant               1     30       5           1
  indépendant            2      9       4           1
```

N = 111

Valeurs propres

```
Eigenvalues
                       Dim.1   Dim.2
Variance               0.052   0.026
% of var.             67.217  32.783
Cumulative % of var.  67.217 100.000
```

Qualité de représentation globale du plan factoriel :

D'après le plan factoriel obtenu et les inerties cumulées, on a une qualité égale à 100%.

<u>Les poids colonnes et lignes :</u>

```
> tab = rowSums(data.ca)
> tab = tab/111
> tab
    employé    étudiant indépendant
  0.5225225   0.3333333   0.1441441
> tab = colSums(data.ca)
> tab = tab/111
> tab
Hebdomadaire       Jamais      Mensuel  Quotidienne
  0.08108108   0.67567568   0.22522523   0.01801802
```

Profils lignes :

```
Rows
             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
employé      |    20.926 |  0.181 32.609  0.815 | -0.086 15.139  0.185 |
étudiant     |    34.535 | -0.320 65.368  0.990 | -0.032  1.299  0.010 |
indépendant  |    22.384 |  0.086  2.023  0.047 |  0.385 83.562  0.953 |
```

Profils colonnes :

```
Columns
             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
Hebdomadaire |    18.665 |  0.454 31.984  0.897 |  0.154  7.562  0.103 |
Jamais       |    14.080 | -0.136 23.883  0.888 | -0.048  6.204  0.112 |
Mensuel      |    18.354 |  0.285 35.076  1.000 |  0.001  0.001  0.000 |
Quotidienne  |    26.745 | -0.513  9.057  0.177 |  1.105 86.233  0.823 |
```

Contribution des profils lignes et colonnes à la construction des axes :

Les profils lignes ou colonnes qui contribuent à la construction d'un axe sont ceux qui ont une contribution absolue supérieure à son poids.

Axe 1 :

PL's		**PC's**	
+	**-**	**+**	**-**
	etudiant	hebdomadaire mensuel	Quotidienne

-L'axe 1 est un axe à effet opposition, il oppose les clients Quotidiens qui sont des étudiants aux clients hebdomadaires et mensuels.

Axe 2 :

PL's		**PC's**	
+	**-**	**+**	**-**
Indépendant		Quotidienne	

-L'axe 2 est un axe à effet taille, il qualifie les clients Quotidiens qu'ils sont des travailleurs indépendants

Qualité de représentation des profils lignes et colonnes par les axes :

Axe 1 :

-Les profils lignes qui sont bien représentées par l'axe 1 sont : étudiant et employé

Car on a :

cos2(étudiant) = 0.815 (proche de 1)

cos2(employé) = 0.99 (proche de 1)

-Les profils colonnes qui sont bien représentées par l'axe 1 sont : hebdomadaire , jamais et mensuel, car on a :

cos2(Hebdomadaire) = 0.897 (proche de 1)

cos2(Jamais) = 0.888 (proche de 1)

cos2(Mensuel) = 1 (exactement égal à 1)

Axe 2 :

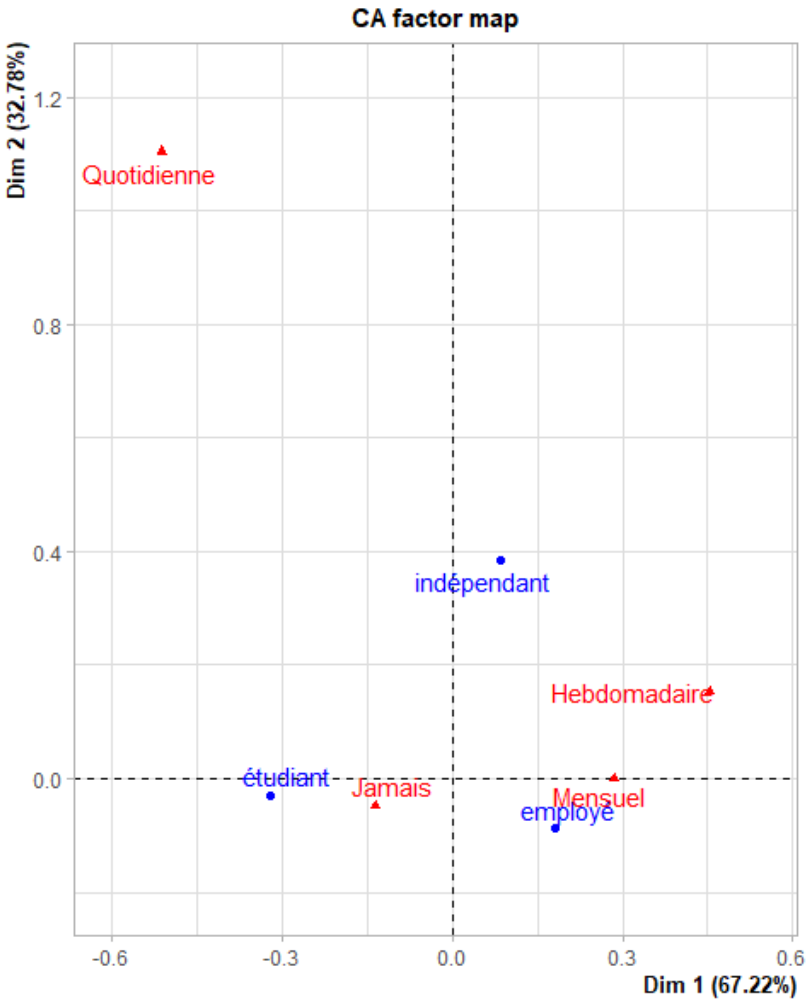
-Les profils lignes qui sont bien représentées par l'axe 2 sont : Indépendant

Car on a : cos2(Indépendant) = 0.953 (proche de 1)

-Les profils colonnes qui sont bien représentées par l'axe 2 sont : Quotidienne

Car on a : cos2(Quotidienne) = 0.823 (proche de 1)

Visualisation des résultats et interprétation :



Remarque :

- D'après l'étude des contributions on voit que la modalité “Quotidienne” contribue à la construction de l’axe 1 mais celle-ci n’est pas bien représentée par cet axe, c'est parce que la modalité “Quotidienne” est une modalité rare, elle a une très petite fréquence (0.018). Donc pour avoir de meilleurs résultats on enlève la modalité “Quotidienne” de la variable visitNo qui est une valeur aberrante et on refait l’étude de l’afc en considérant la catégorie “Quotidienne” comme variable supplémentaire.

- **Nouveau tableau de contingence :**

status	visitNo Hebdomadaire	Jamais	Mensuel
employé	6	36	16
étudiant	1	30	5
indépendant	2	9	4

N = 109

- **On recalcule les poids des modalités :**

```
> tab = rowSums(data.ca)
> tab
    employé    étudiant indépendant
         58          36          15
> tab/109
    employé    étudiant indépendant
  0.5321101   0.3302752   0.1376147
> tab = colSums(data.ca)
> tab
Hebdomadaire       Jamais      Mensuel
           9           75           25
> tab/109
Hebdomadaire       Jamais      Mensuel
  0.08256881  0.68807339  0.22935780
```

Valeurs propres :

```
Eigenvalues
                       Dim.1   Dim.2
Variance               0.051   0.001
% of var.             98.396   1.604
Cumulative % of var.  98.396 100.000
```

Qualité de représentation globale du plan factoriel :

D'après le plan factoriel obtenu et les inerties cumulées, on obtient la même qualité de représentation du plan qui est égale à 100%, de plus on voit qu'avec un seul axe on a une meilleure qualité par rapport à l'étude précédente (de 67% à 98%). En utilisant la méthode des 80%, on décide de ne garder qu'un seul axe factoriel.

profils lignes :

```
Rows
             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
employé      |    11.338 |  0.145 21.777  0.982 | -0.020 25.012  0.018 |
étudiant     |    33.922 | -0.320 66.362  1.000 |  0.004  0.610  0.000 |
indépendant  |     6.682 |  0.210 11.861  0.907 |  0.067 74.377  0.093 |
```

profils colonnes :

```
Columns
             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
Hebdomadaire |    19.113 |  0.475 36.496  0.976 |  0.075 55.248  0.024 |
Jamais       |    15.191 | -0.149 29.698  0.999 |  0.004  1.494  0.001 |
Mensuel      |    17.638 |  0.274 33.806  0.980 | -0.040 43.258  0.020 |
```

```
Supplementary column
                Dim.1   cos2    Dim.2   cos2
Quotidienne   | -0.245  0.038 |  1.230  0.962
```

Contribution des profils lignes et colonnes à la construction des axes :

Les profils lignes ou colonnes qui contribuent à la construction d'un axe sont ceux qui ont une contribution absolue supérieure aux poids.

Axe 1 :

PL's		**PC's**	
+	**-**	**+**	**-**
	etudiant	hebdomadaire mensuel	

-L'axe 1 est un axe à effet taille, il qualifie les clients hebdomadaires et mensuels qui ne sont majoritairement pas des étudiants

Axe 2 :

PL's		**PC's**	
+	**-**	**+**	**-**
Indépendant		hebdomadaire mensuel	

-L'axe 2 est un axe à effet taille, il qualifie les clients hebdomadaires et mensuels qui sont majoritairement indépendants

Qualité de représentation des profils lignes et colonnes par les axes :

Axe 1 :

=> tous les profils lignes (étudiant , employé et indépendant) sont bien représentés par l'axe 1, ils ont tous une contribution relative très proche de 1 .

-On remarque par rapport à l'étude précédente que la modalité "indépendant" n'était pas bien représentée par l'axe 1 mais après suppression de la modalité "Quotidienne", elle est devenue bien représentée par cet axe.

=>Tous les profils colonnes (hebdomadaire , jamais et mensuel) sont bien représentés par l'axe 2 , ils ont tous une contribution relative très proche de 1.

Axe 2 :

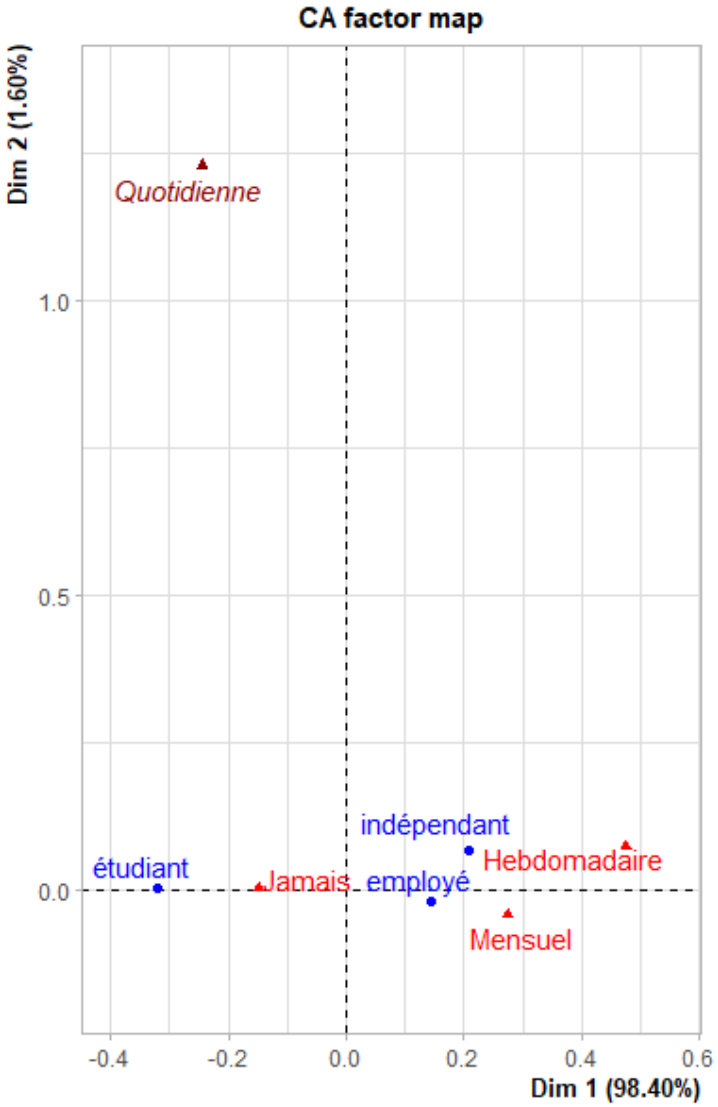
-Aucun profil lignes n'est bien représenté par l'axe 2, ils ont tous une contribution relative proche de 0.

-On a la même chose pour les profils colonnes, aucun profil colonne n'est bien représenté par l'axe 2, ils ont tous une contribution relative proche de 0.

Remarque par rapport à la qualité de représentation des profils lignes et colonnes par les axes :

En comparant les 2 études, on constate que la modalité indépendante qui était bien représentée par l'axe 2 est devenue bien représentée par l'axe 1 et toutes les modalités sont devenues bien représentées par l'axe 1 en revanche aucune ne l'est par l'axe 2. Donc on va faire notre interprétation par rapport à l'axe 1 seulement.

Lors de la première étude, la variable "Quotidienne" avait contribué à la construction de l'axe 1 et a ainsi falsifié l'interprétation par rapport à celui-ci



- D'après la signification de l'axe 1 on a :

les modalités "hebdomadaire" , "mensuel" , et "étudiant" de plus de leurs contribution à la construction de l'axe 1,elles sont bien représentées par ce dernier et donc on peut dire que les clients qui arrivent hebdomadairement et mensuellement ne sont pas des étudiants.

- D'après le graphe on a :

-Les clients qui ont des fréquences de visite hebdomadaires sont généralement des indépendants. On peut ainsi constater que les indépendants sont ceux qui viennent le plus chez Starbucks
-Les clients qui arrivent mensuellement sont généralement des employés.
-Les étudiants ont des visites rares à Starbucks

IV. Conclusion :

Comment est ce que starbuck pourrait tirer profit de ses informations pour améliorer sa stratégie Marketing et de fidélisation de la clientèle ?

1- Etant donné que les jeunes employés dont le salaire ne dépasse pas 50.000 représentent une grande majorité des clients de starbucks, il est préférable de proposer des produits à des prix moins chers pour justement les fidéliser.
2- Etant donné que les jeunes fréquentent plus starbucks que les plus âgés, le marketing peut être ciblé sur les canaux digitaux et les réseaux sociaux

V. Script R :

```
library("ade4")
library('dplyr')
library('FactoMineR')
library('factoextra')

data <- read.csv("Data.csv", header = TRUE, sep = ",")#lecture des données
data
View(data)

# enlever l'id
data = select(data , -1)

#selectionner que les variables à au moins deux modalités
data.new = select(data , -c(10 ,11,12,13,14,15,24,25,26,27,28,29,30,31))

tab = summary(data.new$loyal) #Code pour avoir les occurences de chaque modalité des variables
freq= tab/113 #pour avoir les fréquences
freq

#selectionner les variables pour afcm
data.new =select(data.new , c('gender', 'age', 'status', 'income' ,'visitNo'))

#Supprimer quelques modalités non pertinentes
data.new<-data.new[!(data.new$status==3),]
data.new <- data.new[ data.new$status !=3, , drop=FALSE];
#data.new$status <- factor(data.new$status);
data.new$income[data.new$income == 4] <- 3
data.new <- data.new[ data.new$income !=4, , drop=FALSE];
#data.new$income <- factor(data.new$income);

#renommer les variables

data.new$age[data.new$age == 0]<- "<20"
data.new$age[data.new$age == 1]<- "20<age<29"
data.new$age[data.new$age == 2]<- "30<age<39"
data.new$age[data.new$age == 3]<- ">40 ans"

data.new$status[data.new$status == 0] <- "étudiant"
data.new$status[data.new$status == 1]<- "indépendant"
data.new$status[data.new$status == 2] <- "employé"

data.new$gender[data.new$gender == 0] <- "Homme"
```

```
data.new$gender[data.new$gender == 1]<- "Femme"

data.new$income[data.new$income == 0] <- "salaire<25000"
data.new$income[data.new$income == 1] <- " 25000 < salaire < 50000"
data.new$income[data.new$income == 2] <- " 50000 < salaire < 100000"
data.new$income[data.new$income == 3]<- " 100000 < salaire"

data.new$visitNo[data.new$visitNo == 0] <- "Quotidienne"
data.new$visitNo[data.new$visitNo== 1] <- "Hebdomadaire"
data.new$visitNo[data.new$visitNo == 2] <- "Mensuel"
data.new$visitNo[data.new$visitNo == 3] <- "Jamais"

data.new$spendPurchase[data.new$spendPurchase == 0] <- "depense=Zero"
data.new$spendPurchase[data.new$spendPurchase == 1]<- "depense<20"
data.new$spendPurchase[data.new$spendPurchase == 2]<- "20<depense<40"
data.new$spendPurchase[data.new$spendPurchase == 3] <- "depense>40"

data.new$location[data.new$location == 0] <- "R<1KM"
data.new$location[data.new$location == 1]<- "2 et 3 Km"
data.new$location[data.new$location == 2] <- "R >3km"

#tableau disjonctif
data.disj = acm.disjonctif(data.new)
View(data.disj)

#transformer les valeurs des variables de chaînes de caractère à facteur
i=0
while(i < ncol(data.new)){
  i=i+1
  data.new[,i] = as.factor(data.new[,i])
}

#Faire l'afcm
mca = MCA(data.new)

#corriger les inerties
vp = mca$eig
vp = vp[,1]
vp = vp [vp>1/5]
vpa = 5/4 * (vp - 1/5)
I = vpa / sum(vpa) * 100

#visualisations
fviz_mca_var(mca, repel = TRUE , col.var = "Blue" )

fviz_mca_var(mca, col.var="cos2",
```

```
        gradient.cols = c( "#ffba08","#d00000","#03071e"),
        repel = TRUE # Avoid text overlapping
        )
fviz_contrib(acm, choice ="var" , axes=1 )
fviz_contrib(acm, choice ="var" , axes=2 )
plot (data.new$visitNo ,main ="visitNo" ,col ="#0892A5")
plotellipses(mca)
fviz_mca_ind (mca,
        label = "none", # masquer le texte des individus
        habillage = "visitNo", # colorer par groupes
        addEllipses = TRUE,
        ellipse.type = "confidence",
         ggtheme = theme_minimal ())
fviz_mca_ind (mca,
         label = "none", # masquer le texte des individus
         habillage = "visitNo", # colorer par groupes
         addEllipses = TRUE,
         ellipse.type = "confidence",
         ggtheme = theme_minimal ())

barplot(freq , main ="loyal", col ="#0892A5")
fviz_mca_biplot (mca, repel = TRUE,
            col.var = "Purple" , ind.var="Orange")

qplot(c(1:11), l) +
 geom_col()+
 geom_line() +
 geom_point(size=4)+
 xlab("Principal Component") +
 ylab("Variance Explained") +
 ggtitle("Scree Plot") +
 ylim(0, 50)

 # effectuer l'AC
data.ca <- read.csv("Data.csv", header = TRUE, sep = ",")#lecture des données
data.ca = select(data.ca , c('status','visitNo')) #sélection des variables
data.ca = table(data.ca) #générer le tableau de contingence

ca = CA (data.ca , col.sup = 4) #AFC
```