Analyse du comportement bancaire de clients

Makhatch ABDULVAGABOV, Florent VALBON, Samuel ASSARAF

Janvier 2020

Table des matières

1 Introduction

Le jeu de données étudié permet d'analyser le comportement bancaire des 500 clients. Les données sont issues d'une enquête réalisée régulièrement par une banque pour créer de nouveaux produits afin de fidéliser les clients.
Il s'agira de proposer une typologie des individus et d'identifier qui sont, en général, les clients qui gèrent leur compte d'une certaine façon.

2 Présentation des données

```
rm(list=ls())
library("gridExtra")
library("cowplot")
library("corrplot")
library(knitr)
library(kableExtra)
library(FactoMineR)
library(factoextra)
library(tidyverse)
```

```r
library(rmarkdown)
library(gridExtra)
library(grid)
library(png)
library(downloader)
#library(grDevices)
#setwd("~/Desktop")
cbg <- read.table(file = "cbg500.txt", head = T, dec = ",")
head(cbg)
```

```
##     solde mdecouv ncompte memprunt   mdepot mretrait nbenf age
## 1  245.00 1139.67       0   129.57    31.50     0.00     0  22
## 2 2326.07    0.00       2  3810.98 63516.26  2330.18     0  45
## 3  188.41 1503.14       1     0.00    60.98   785.39     0  19
## 4 1256.25  227.96       9 32012.20  2439.03 14246.30     0  62
## 5  946.65  305.66       9 17225.61 11432.93  9291.94     0  36
## 6 1047.41  487.69       7 30487.80  9527.44  5688.58     0  31
##                  csp code
## 1              autre    6
## 2 artisan-commercant    1
## 3              autre    6
## 4           retraite    5
## 5            ouvrier    4
## 6            ouvrier    4
```

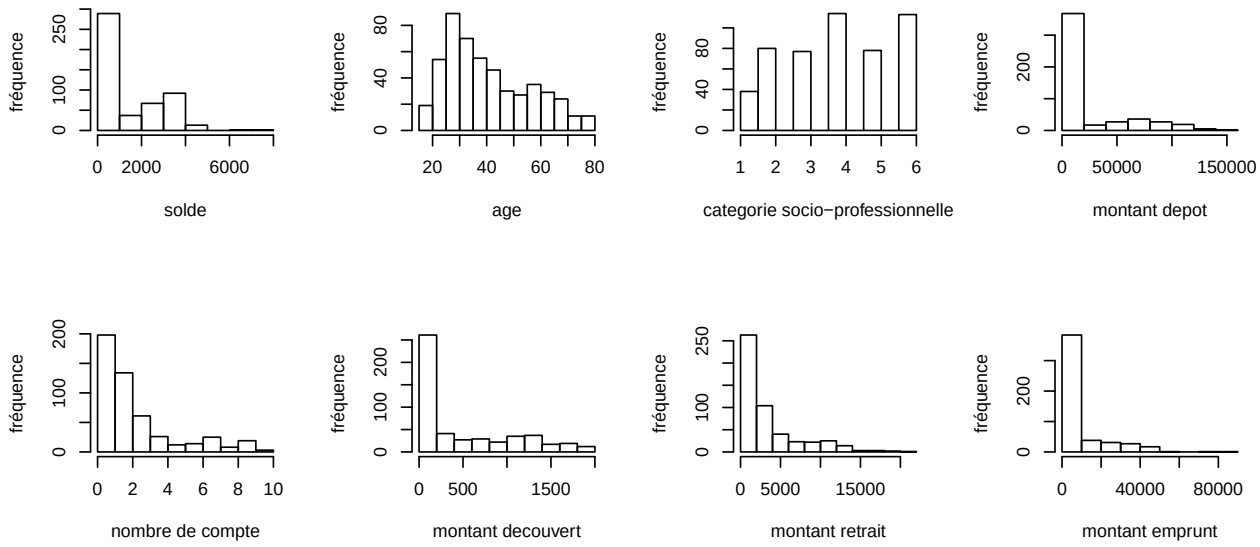
Le jeu de données `cbg` contient 500 individus (clients) sur lesquels on mesure 10 variables :

- **solde** : Solde moyen du compte courant sur les 12 derniers mois (en euros) (quantitative)
- **mdecouv** : Montant cumulé des découverts sur le compte courant durant les 12 derniers mois (en euros) (quantitative)
- **ncompte** : Nombre de comptes utilisés en plus du compte courant (par exemple les livrets ...) (quantitative)
- **memprunt** : Montant total des emprunts effectués sur les trois dernières années (en euros) (quantitative)
- **mdepot** : Montant total des versements effectuées sur le livret d'épargne lors des 5 dernières années (en euro) (quantitative)
- **mretrait** : montant total des retraits effectuées sur le livret d'épargne sur les 12 derniers mois (en euros) (quantitative)
- **nbenf** : nombre d'enfants de moins de 18 ans (quantitative)
- **age** : age du client enquêté (quantitative)
- **csp** : catégorie socio-professionnelle du client (qualitative)
- **code** : codification de la catégorie socio-professionnelle du client (qualitative)

    1. artisant-commerçant
    2. cadre
    3. employé
    4. ouvrier
    5. retraité
    6. autre

```r
#names(cbg)
#str(cbg)
par(mfrow=c(2,4))
hist(cbg$solde,main=" ",xlab = "solde", ylab = "fréquence",)
hist(cbg$age,main=" ",xlab = "age", ylab = "fréquence",)
```

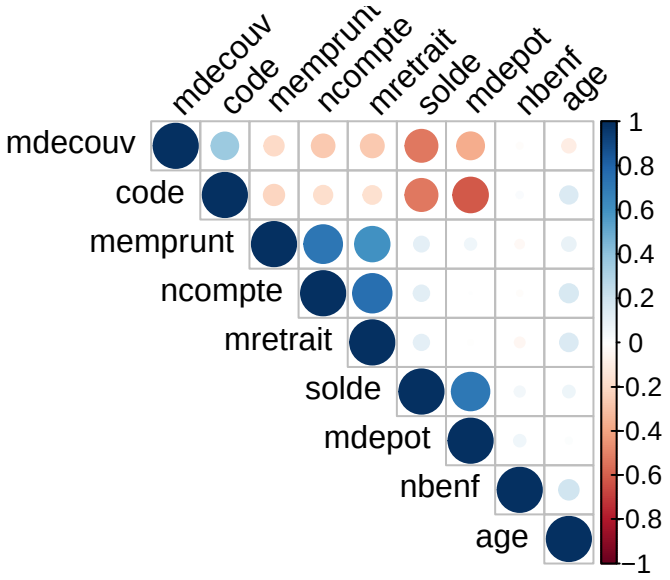
```
hist(cbg$code,main=" ",xlab = "categorie socio-professionnelle", ylab = "fréquence",)
hist(cbg$mdepot,main=" ",xlab = "montant depot", ylab = "fréquence",)
hist(cbg$ncompte,main=" ",xlab = "nombre de compte", ylab = "fréquence",)
hist(cbg$mdecouv,main=" ",xlab = "montant decouvert", ylab = "fréquence",)
hist(cbg$mretrait,main=" ",xlab = "montant retrait", ylab = "fréquence",)
hist(cbg$memprunt,main=" ",xlab = "montant emprunt", ylab = "fréquence",)
```



L'échantillon de 500 clients est majoritairement composé de clients dont l'activité bancaire est faible : Cette majorité emprunte peu, retire et dépose moins d'argent, et dispose d'un faible nombre de compte. Le troisième quartile pour la variable **nbenf** est de 1 et la médiane vaut 0. Ainsi, 75% des clients de la banque ont au plus un enfant et la moitié n'en n'ont aucun. En moyenne l'âge des clients est de 37 ans et la catégorie socio-professionnelle la plus représentée est ouvrier.

3 Analyse des corrélations

```
corrplot(cor(cbg[-9]), type="upper", order="hclust", tl.col="black", tl.srt=45)
```



Naturellement, il apparait que les montants d'emprunt, de retrait et de compte sont corrélés positivement deux à deux, et le montant du dépôt l'est avec le solde. Aussi, le montant du découvert est corrélé négativement avec le solde. Ces corrélations (assez forte : strictement supérieur à 0.7 en valeur absolue) se conçoivent logiquement.

4 Analyse de l'impact de la catégorie socio-professionnelle et de l'âge sur le comportement bancaire des clients

Dans cette partie, nous analyserons le comportement bancaire des clients par deux variables : **solde** et **memprunt**. Nous étudierons les impacts qu'ont la catégorie socio-professionnelle et l'âge sur ces variables.

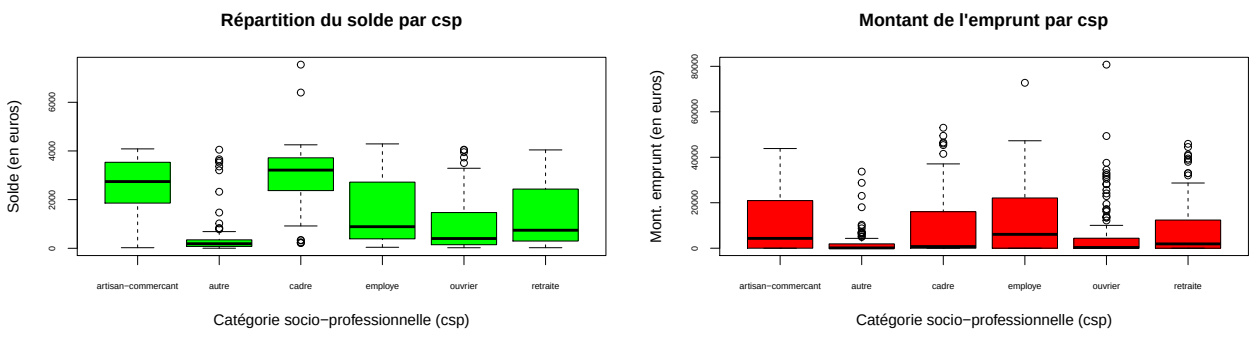
4.1 Comportement en fonction de la catégorie socio-professionnelle

```
cbg500 = cbg;
```

Les boxplots nous montrent les distributions des individus en fonction de leurs catégories socio-professionnelle.

```
par(mfrow=c(1,2))
boxplot(solde~csp, data = cbg500, "Solde en fonction de la csp (euros)", xlab = "Catégorie socio-profes

boxplot(memprunt~csp, data = cbg500, col="red", main = "Montant de l'emprunt par csp", xlab = "Catégori
```

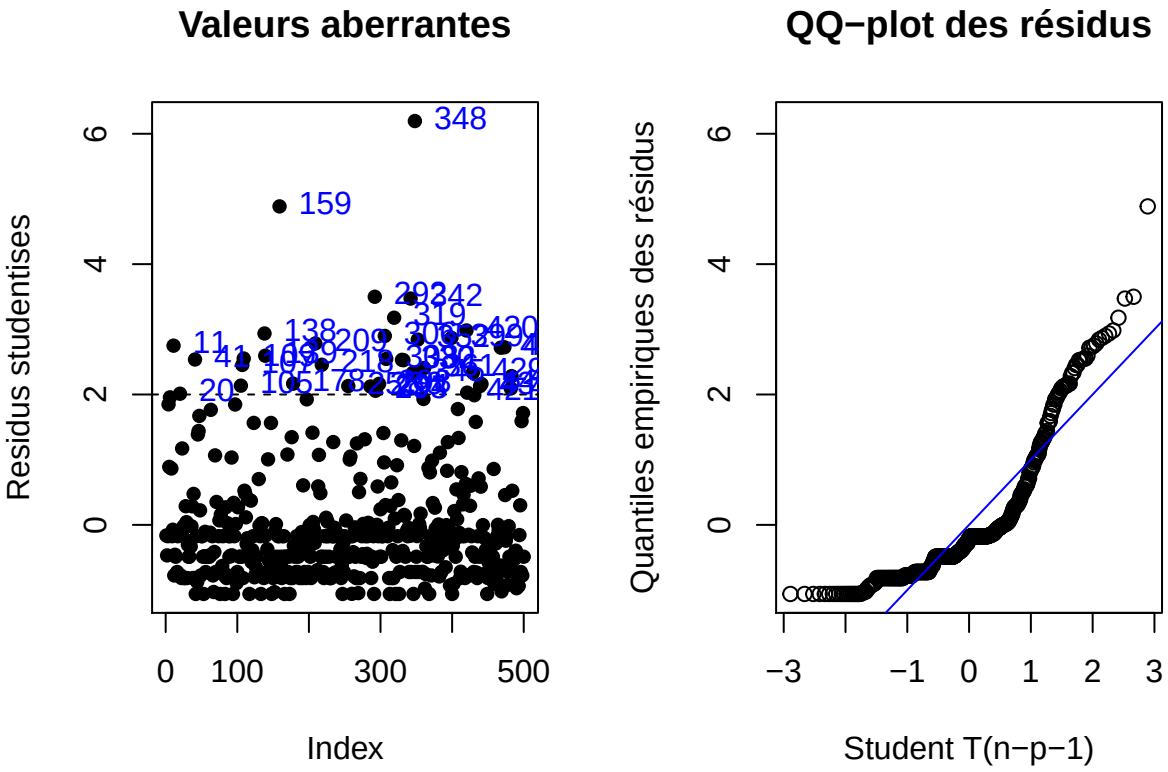


Les boxplots sont très différents deux à deux pour la variable **solde**. **Dans ce cas, l'impact de la catégorie socio-professionnelle est clair : Les cadres ont le solde moyen le plus élevé parmi toutes les catégories socio-professionnelles tandis que les "autres" catégories auront le solde le plus faible (il est fort probable que cette catégorie inclue les personnes sans activité/étudiantes).** En revanche, les boxplots de la variable **memprunt** ne sont que légèrement différents. Nous allons donc procéder à une **ANOVA** afin de voir si ces différences sont significatives ou non.
Vérifions dans un premier temps que les hypothèses sont vérifiées :

```
mod1 = lm(memprunt~csp,data = cbg500);
```

```
par(mfrow=c(1,2))
residus <- rstudent(mod1)
n <- length(cbg500$mretrait)
plot(1:n, residus, pch = 16, xlab = 'Index', ylab = 'Residus studentises',
     main = 'Valeurs aberrantes')
abline(-2, 0, lty = 2)
abline(2, 0, lty = 2)
IDval.ab <- (1:n)[abs(residus)>2]
text(IDval.ab, residus[IDval.ab], IDval.ab, pos = 4, col = 'blue')


quant.t = qt((1:500)/n,n-3);
plot(quant.t, sort(residus), xlab = 'Student T(n-p-1)',
     ylab = 'Quantiles empiriques des résidus', main = 'QQ-plot des résidus')
abline(0, 1, col ='blue')
```



Peu d'observations sont en dehors de l'intervalle [-2,2] par rapport au nombre total de données. Nous ne supprimons toutefois pas celles qui sont en-dehors de cet intervalle. D'autre part, l'alignement des points sur la première bissectrice est plus ou moins vérifiée, ce qui confirme l'hypothèse selon laquelle les résidus théoriques du modèle linéaire gaussien associé suivent une loi normale centrée est bien vérifiée. Enfin, il

n'apparait pas de structures dans les résidus.
Le modèle **ANOVA** est donc valide :

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: memprunt
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## csp          5 6.8428e+09 1368567285  8.6924 6.523e-08 ***
## Residuals 494 7.7777e+10  157443551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-value du test de Fisher est très faible (**6.523e-08**), on en conclut que **la catégorie socio-professionnelle a effectivement un impact sur le montant total des emprunts effectués sur les 3 dernières années.**

On montrera de même que l'ensemble des modèles **ANOVA** utilisés pour cette étude sont valides.

4.2 Comportement en fonction de l'âge

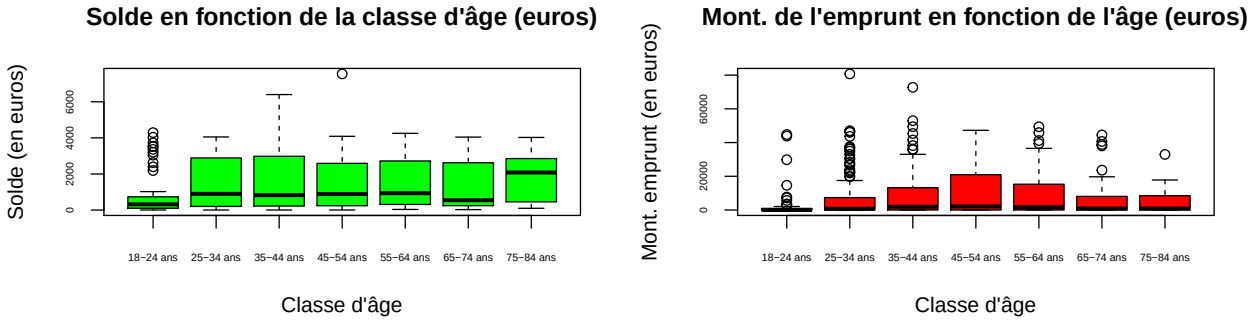
Les boxplots suivants modélisent la distribution des clients en fonction de leurs âges :

```
par(mfrow=c(1,2))
classe_age = c();
for(i in 1:500){
  if(cbg500$age[i]<=24){
    classe_age[i] = "18-24 ans";
  }
   if((cbg500$age[i]>24)&&(cbg500$age[i]<=34)){
    classe_age[i] = "25-34 ans";
   }
  if((cbg500$age[i]>34)&&(cbg500$age[i]<=44)){
    classe_age[i] = "35-44 ans";
  }
  if((cbg500$age[i]>44)&&(cbg500$age[i]<=54)){
    classe_age[i] = "45-54 ans";
  }
  if((cbg500$age[i]>54)&&(cbg500$age[i]<=64)){
    classe_age[i] = "55-64 ans";
  }
  if((cbg500$age[i]>64)&&(cbg500$age[i]<=74)){
    classe_age[i] = "65-74 ans";
  }
  if((cbg500$age[i]>74)&&(cbg500$age[i]<=84)){
    classe_age[i] = "75-84 ans";
  }
}

cbg500_ca = cbind(cbg500,classe_age);

boxplot(solde~classe_age, data = cbg500_ca, main = "Solde en fonction de la classe d'âge (euros)", xlab

boxplot(memprunt~classe_age, main = "Mont. de l'emprunt en fonction de l'âge (euros)", xlab = "Classe d
```



Au vu des données, il nous a semblé pertinent de regrouper les individus par tranches d'âge. Observons également le nombre de clients pour chaque classe d'âge.

```
summary(cbg500_ca$classe_age)
```

```
## 18-24 ans 25-34 ans 35-44 ans 45-54 ans 55-64 ans 65-74 ans 75-84 ans
##        62       162        95        69        62        39        11
```

Pour les deux variables, les boxplots sont tous légèrement différents. Une **ANOVA** permettra de voir si ces différences sont significatives ou non. Pour la variable **solde**, il n'y a que 11 individus de plus de 74 ans. Ce n'est pas suffisant pour établir une règle générale. Nous ne nous intéresserons qu'aux individus âgés de 25 à 74 ans, car de plus, contrairement au plus de 74 ans, les boxplots associés ne diffèrent légèrement que sur cet intervalle d'âge.

```
mod3 = lm(solde[(24<age)&(age<75)]~classe_age[(24<age)&(age<75)],data = cbg500_ca);
```

```
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: solde[(24 < age) & (age < 75)]
##                                      Df    Sum Sq Mean Sq F value Pr(>F)
## classe_age[(24 < age) & (age < 75)]   4   1553258  388314  0.1827 0.9473
## Residuals                           422 896969679 2125521
```

La p-value n'est pas faible (**94%**). Cela montre qu'**entre 25 et 74 ans l'âge n'a pas d'impact significatif sur le solde moyen du compte courant sur 12 mois.**

En revanche, la p-value du test de Fisher de **3,3%** dans le modèle **ANOVA** modélisant l'effet du facteur **age** sur **memprunt** montre qu'**avec une incertitude de 5%, l'âge a un impact significatif sur le montant total des emprunts en 3 ans.**

```
mod4 = lm(memprunt~age,data = cbg500);
```

```
anova(mod4)
```

```
## Analysis of Variance Table
##
## Response: memprunt
##            Df     Sum Sq   Mean Sq F value Pr(>F)
## age         1 7.6883e+08 768829710  4.5662 0.0331 *
## Residuals 498 8.3851e+10 168375745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5 Réalisation de l'ACP

Pour réaliser l'analyse en composantes principales de ce jeux de données, les variables **csp** et **code** ne seront pas prises en compte (déclarées qualitatives supplémentaires).

```
res.pca<-PCA(cbg, scale.unit = TRUE,  quali.sup=9:10, graph=F)
```

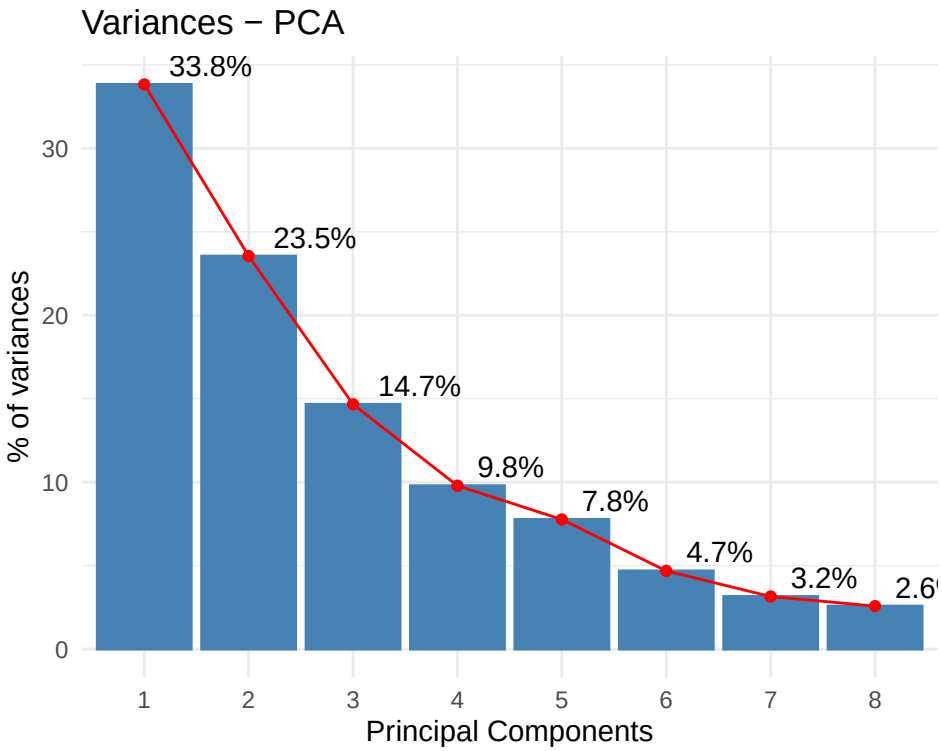
5.1 Choix de la dimension

Pour choisir le nombre de composantes principales, on peut s'appuyer sur le tableau et les figures suivantes.

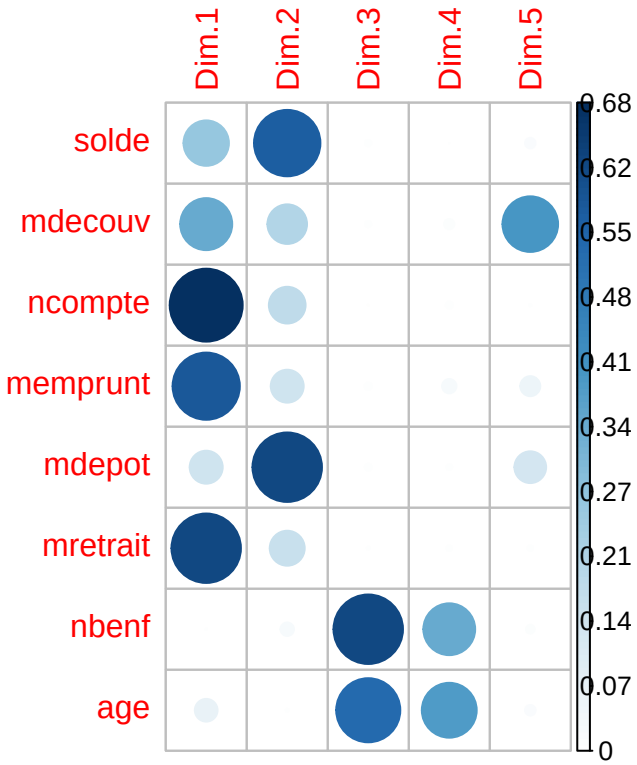
```
eig.val <- get_eigenvalue(res.pca)
kable(eig.val)
```

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.7059279	33.824098	33.82410
Dim.2	1.8833722	23.542153	57.36625
Dim.3	1.1730582	14.663227	72.02948
Dim.4	0.7830071	9.787589	81.81707
Dim.5	0.6214060	7.767574	89.58464
Dim.6	0.3748707	4.685884	94.27053
Dim.7	0.2521917	3.152396	97.42292
Dim.8	0.2061662	2.577078	100.00000

```
p <- fviz_eig(res.pca, addlabels=TRUE, hjust = -0.3, linecolor ="red") + theme_minimal() + labs(title =
print(p)
```




```
var <- get_pca_var(res.pca)
corrplot(var$cos2, is.corr=FALSE)
```

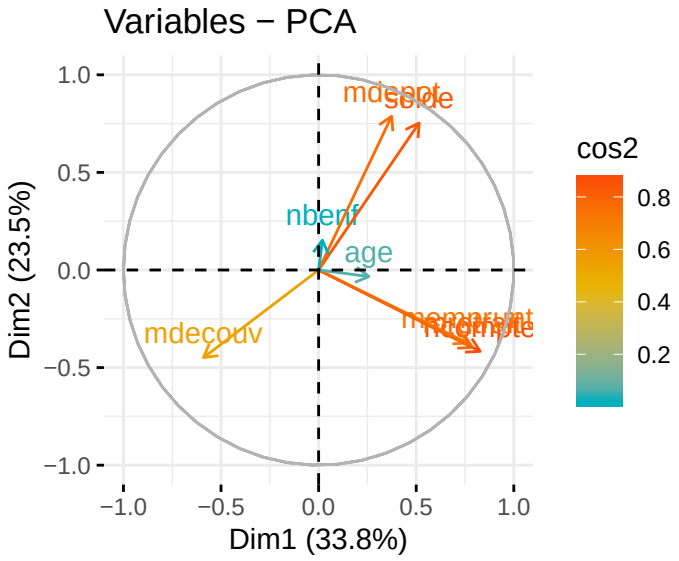


Le tableau indique qu'avec 4 dimensions, presque 82% de la variance est conservé. De même, le graphe indique que la dimension 5 n'apporte que 7.8% de variance. De plus, la figure des contribution des variables aux dimensions montre que toutes les variables contribuent principalement aux 4 premières dimensions (sauf **mdecouv** qui contribue légèrement à la dimension 5). On peut donc judicieusement s'arrêter à 4 dimensions.

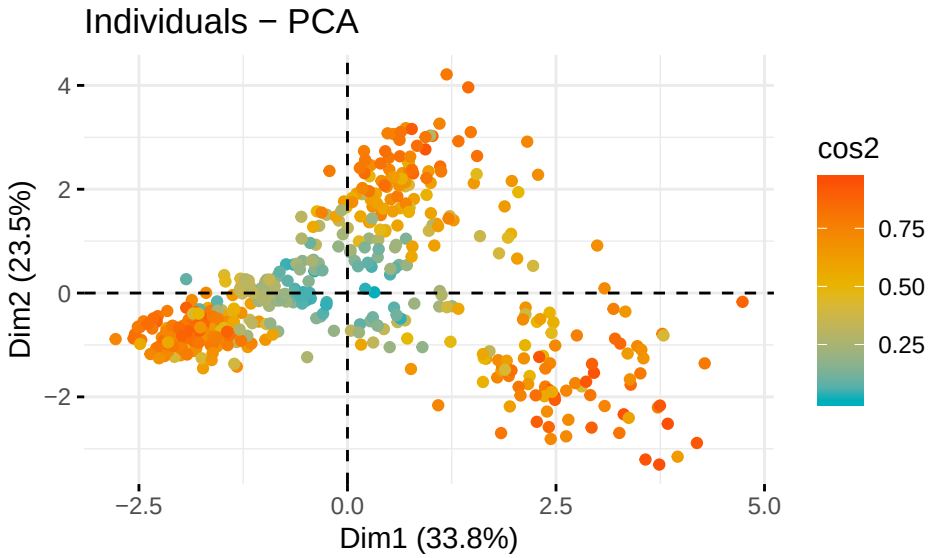
5.2 Cercles de corrélation et nuages de points

5.2.1 Premier plan

```
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             #repel = TRUE # Évite le chevauchement de texte
             ) #-> p1
```




```
fviz_pca_ind (res.pca, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             geom.ind = "point", # Montre les points seulement (mais pas le "text")
             )# -> p2
```



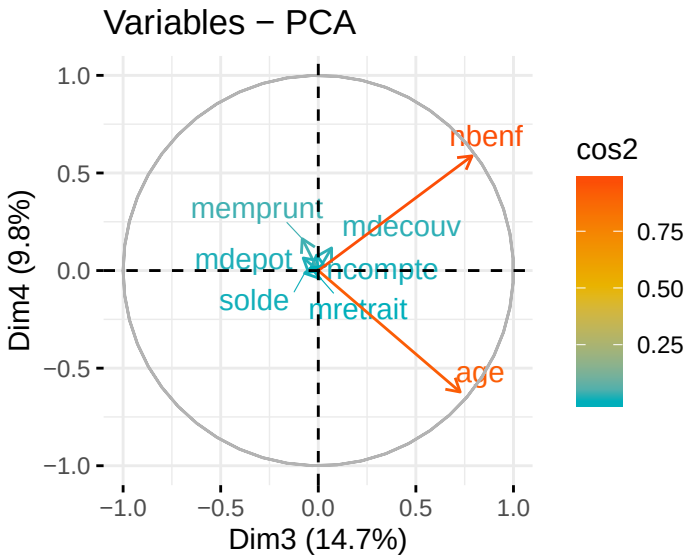

```
#grid.arrange(p1, p2, ncol = 2)
```

Dans le premier plan (dimension 1-2), le cercle de corrélations montre que les variables mdepot et solde sont corrélées entre elles. Ce résultat est assez intuitif, car plus un client dépose de l'argent, plus le solde sera important. Il est aussi naturel de constater que la variable mdecouv est corrélée négativement à ces deux variables, car le fait d'être à découvert est lié au fait d'avoir un solde faible. Les variables mretrait, memprunt et ncompte sont très corrélées entre elles. On pourrait supposer qu'une personne qui retire beaucoup d'argent est susceptible d'avoir plusieurs comptes et faire des emprunts. Ce cercle relève aussi l'absence de corrélation entre le premier groupe de variables correlés et les secondes (retrait, compte, emprunt). cela se visualise par une orthogonalité forte. Il faut néanmoins avoir en tête que ces caractéristiques sont vrais sur ces 2 premières composantes principales.
En outre, presque tous les vecteurs sont bien représentés sur ce plan (sauf le nombre d'enfants et l'âge). Cela se vérifie par le fait qu'ils sont assez proches de la circonférence du cercle.

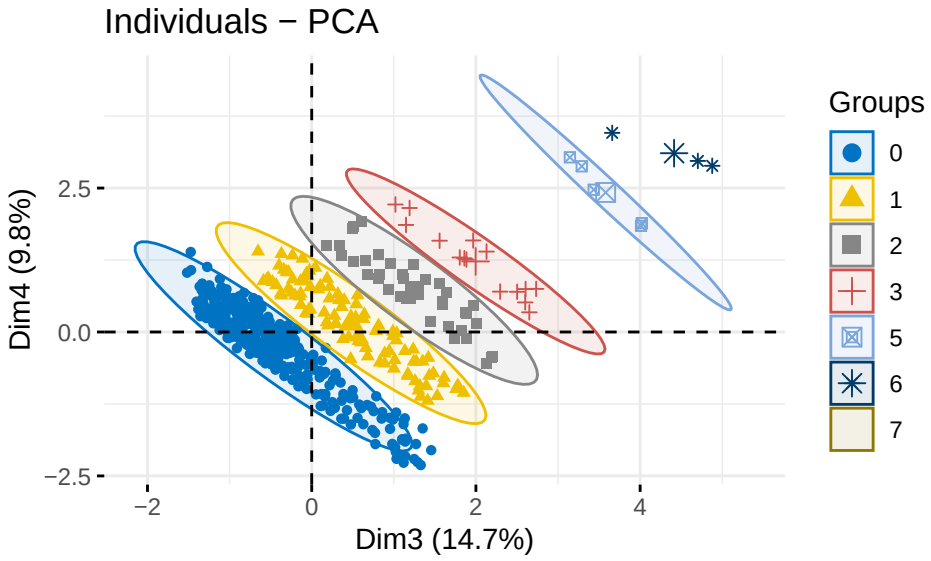
Le nuage de points montre que l'on pourrait séparer les individus en trois groupes dans ce plan. Un premier groupe composé des clients souvent à découvert et n'ayant pas un solde élevé, un deuxième avec les clients présentants l'inverse de ces deux caractéristiques, et un troisième constitué des clients faisant beaucoup d'emprunts et de prélèvements et ayant plusieurs comptes.

5.2.2 Deuxième plan

```
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Évite le chevauchement de texte
             ,axes= 3:4)
```




```
cbg$nbenf <- as.character(cbg$nbenf)
fviz_pca_ind (res.pca, #col.ind = "cos2",
             geom.ind = "point", # Montre les points seulement (mais pas le "text")
             col.ind = cbg$nbenf, # colorer by groups
             palette = "jco",#c("#00AFBB", "#E7B800", "#FC4E07", "#985717", "#997A8D", "#149414","#008E
             #gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07", "#985717", "#997A8D", "#149414"),
             addEllipses = T,
             legend.title = "Groups",
             #repel = TRUE,  Évite le chevauchement de texte
             axes = 3:4
             )
```



Dans le second plan (dimension 3-4), seuls les vecteurs **nbenf**, **age** sont bien représentés sur le cercle des corrélation.
Sur le nuage de points on distingue clairement plusieurs groupes d'individus (entourés par les ellipses). En superposant ce résultat avec le cercle des corrélations, il apparait que la direction dans laquelle les groupes sont divisés est la même que la direction du vecteur **nbenf** (nombre d'enfants). Ce qui explique aussi le caractère discontinu.
À l'intérieur de chaque groupe (ellipse), les points ont une distribution qui parait être centrée sur le centre de l'ellipse. De même que pour le cercle des corrélations, on constate que la direction dans laquelle les points sont distribués est la même que la direction du vecteur **age** (on pourrait par exemple supposer que cette distribution suit une loi gaussienne).

6 Conclusion

L'étude relève immédiatement des corrélations fortes entre les variables. Pour la banque, cela va engendrer une réelle différence de gestion des clients en fonction des caractéristiques qu'ils présentent comme leur catégorie socio-professionnelle ou leur âge. Cela va s'illustrer par une offre de produits adaptés à leur comportement. Par exemple, une personne susceptible de faire beaucoup d'emprunts se verra proposer un prêt. L'ACP quand à elle met en évidence des similarités (solde et dépôt par exemple) et des oppositions entre les variables (solde et retrait par exemple).
Cependant, vu le faible nombre de clients âgés de plus de 74 ans dans les données, nous n'avons pas pu tirer de conclusions sur leur comportement bancaire.
De plus, il serait intéressant d'étudier l'éventuelle interaction entre les facteurs âge et catégorie socio-professionnelle.