Examen 2019: Data Mining

Christophe Tiet

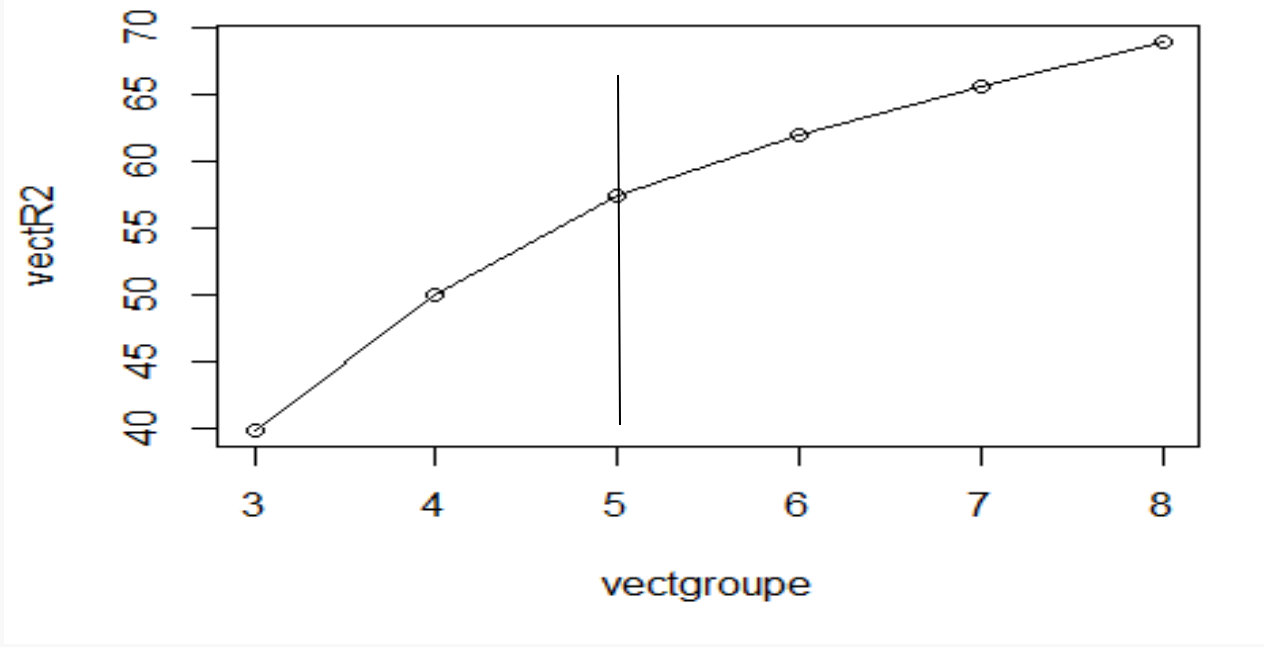
Exercice I

1. Classification des médecins par agrégation autour des moyennes mobiles

On standardise les données car elles n'ont pas le même ordre de grandeur (pourcentage, prix en franc, nombre de visite, age)

On commence par choisir le nombre de groupes pour notre classification.

```
#k=3 R2=39,8%
#k=4 R2=50.0%
#k=5 R2=57.4%
#k=6 R2=61.9%
#k=7 R2=65.5%
#k=8 R2=68.9%
```



La rupture de pente du R2 se produit à k=5 (après, le R2 augmente moins vite). On retient donc **k=5 groupes avec un R2 global de 57.4%.**

```
#Calcul des R2 par variable


##                Eta2      P-value
## AGE       0.6896577 1.161555e-12
## HONPPAT   0.6775674 3.085430e-12
## SHAREFRE  0.6415040 4.613720e-11
## ANCINS99  0.5799749 2.578843e-09
## VISITSHA  0.5532206 1.228804e-08
## MTH70     0.4766946 6.524514e-07
## CONSUPPA  0.3967413 2.222071e-05
```

Toutes les variables ont un R2>0.5 excepté MTH70 et CONSUPPA. Ces deux variables ne peuvent donc pas être commentées et je décide donc de les supprimer de notre jeu de données.

```
#Calcul des moyennes par groupe pour la typologie

##         ANCINS99 SHAREFRE VISITSHA HONPPAT   AGE
## moygr1    71.10     20.41      0.20   401.14 39.95     vert: + de la moyenne
## moygr2   268.44     20.81      0.15   555.28 55.22     rouge - de la moyenne
## moygr3   211.67     17.64      0.28   355.28 50.73
## moygr4   240.50     26.57      0.44   667.72 56.00
## moygr5   134.67     58.03      0.14   458.06 46.00
## moyech   172.32     22.80      0.25   463.17 48.33  ➔   MOYENNE
```

<u>Groupe 1: médecins remplacants</u>

-Ancienneté du medecin très faible

-pourcentage de patients ne payant pas les frais médicaux dans la moyenne

-proportion de visites à domicile dans la moyenne

-honoraire moyen par patient faible

-age du médecin tres faible

<u>Groupe 2: médecins specialistes du privé</u>

-Ancienneté du medecin très elevée

-pourcentage de patients ne payant pas les frais médicaux faibles

-proportion de visites à domicile faible

-honoraire moyen par patient élevé

-age du médecin élevé

Groupe 3: médecins généralistes

-Ancienneté du médecin elevée

-pourcentage de patients ne payant pas les frais médicaux modéramment faible

-proportion de visites à domicile dans la moyenne

-honoraire moyen par patient faible

-age du medecin un peu plus élevé que la moyenne

Groupe 4: practiciens à domicile pour personne à mobilité reduite (personnes agées)

-Ancienneté du medecin elevée

-pourcentage de patients ne payant pas les frais médicaux modéremment élevé

-proportion de visites à domicile élevé

-honoraire moyen par patient élevé

-age du medecin élevé

Groupe 5: médecins du secteur public

-Ancienneté du medecin faible

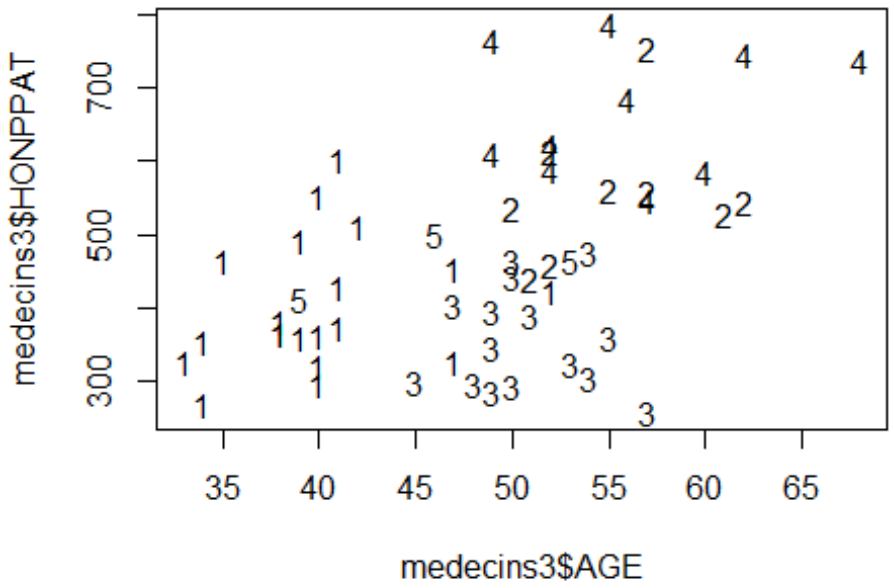
-pourcentage de patients ne payant pas les frais médicaux élevé

-proportion de visites à domicile faible

-honoraire moyen par patient dans la moyenne

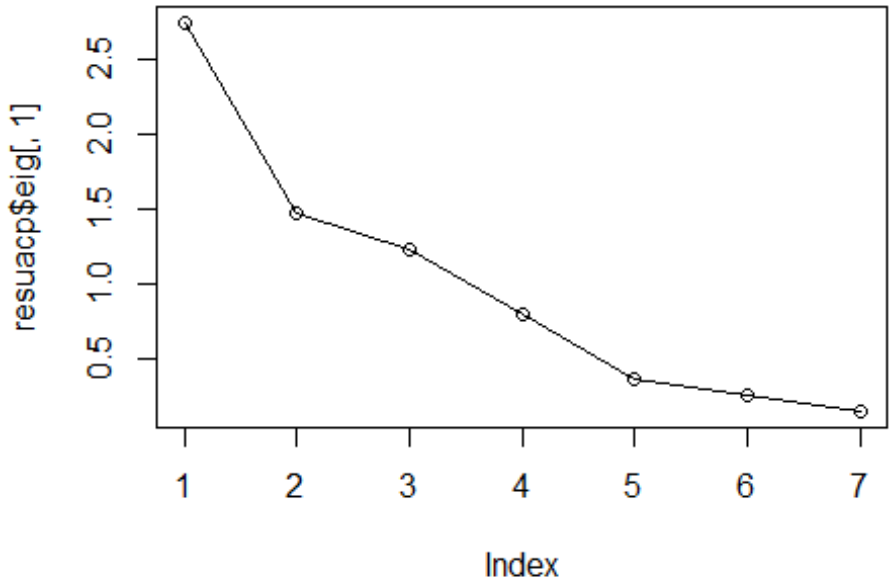
-age du médecin faible

```
Graphique représentant les groupes (nuage de points sur les 2 variables de pl
us fort R2). Comme on représente une information incomplète (2 variables), le
s groupes ne sont pas complètement séparés.
```

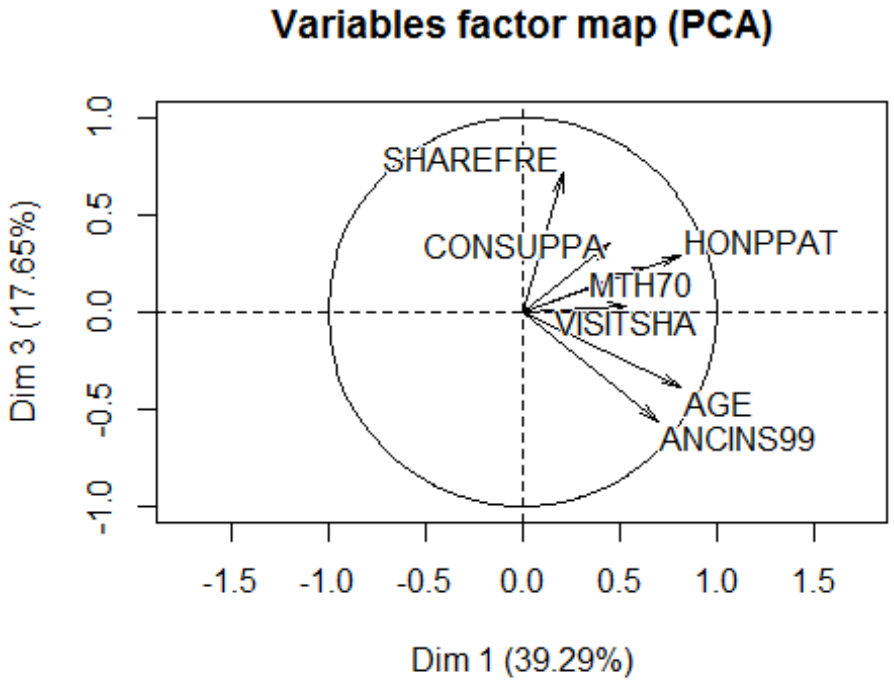


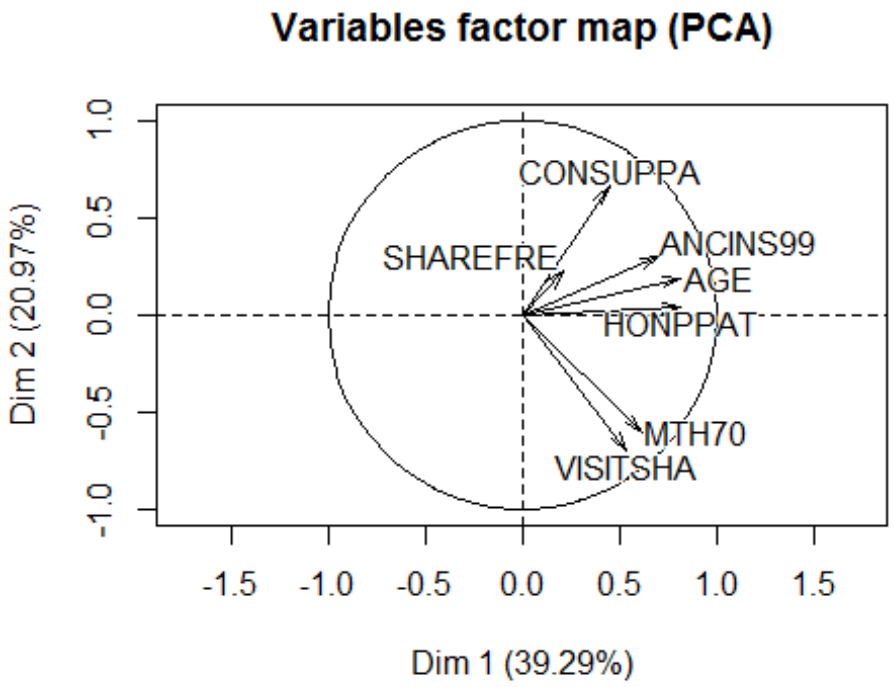
2. Classification ascendante hiérarchique sur composantes principales

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1  2.7502233              39.288904                           39.28890
## comp 2  1.4677402              20.967718                           60.25662
## comp 3  1.2353282              17.647545                           77.90417
## comp 4  0.7927836              11.325479                           89.22965
## comp 5  0.3587763               5.125375                           94.35502
## comp 6  0.2501834               3.574049                           97.92907
## comp 7  0.1449651               2.070930                          100.00000
```



On décide de retenir **les 3 premières composantes principales** (valeur propre >1) afin de garder 77.9% de part d'inertie expliquée malgré la 1ere rupture de pente qui se situe à la deuxième composante principale.

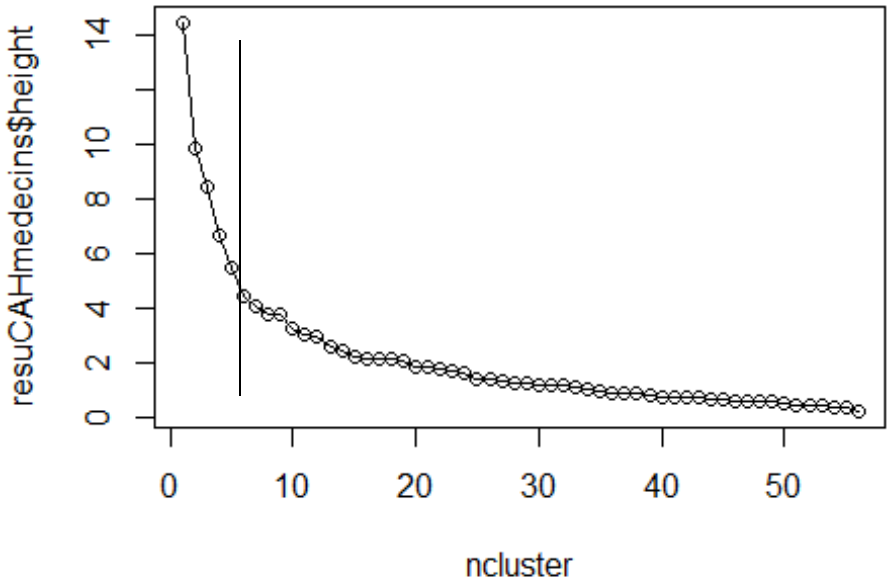




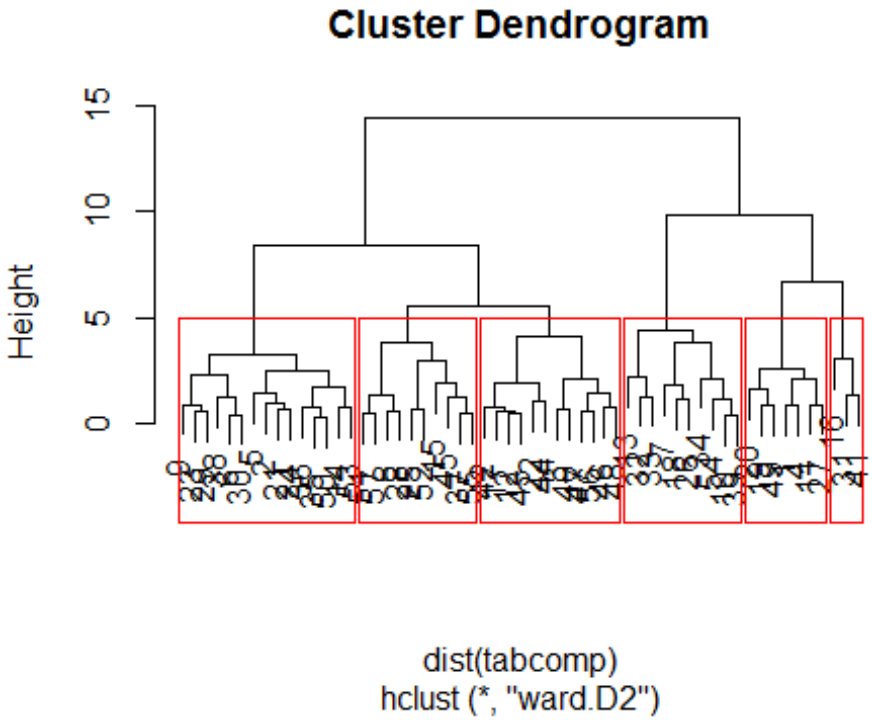
La 1ère CP est positivement corrélée avec l'age, l'ancienneté du médecin et les honoraires par patient. Plus cette CP est élevée et plus le médecin coûte cher et a d'expérience.

La 2ème CP est positivement corrélée avec le nombre de consultations par patient et négativement corrélée avec le taux de patients de plus de 70 ans ainsi qu'avec le pourcentage de visites à domicile. Si cette CP est elevée alors le médecin a un taux de patients de plus de 70 ans faible et se déplace peu à domicile mais le nombre de consultations par patient est élevé.

La 3ème CP a une corrélation positive avec le taux de patients qui ne payent pas de frais médicaux. Plus cette CP est élevée et plus le taux de patients ne payant pas de frais médicaux est élevé.



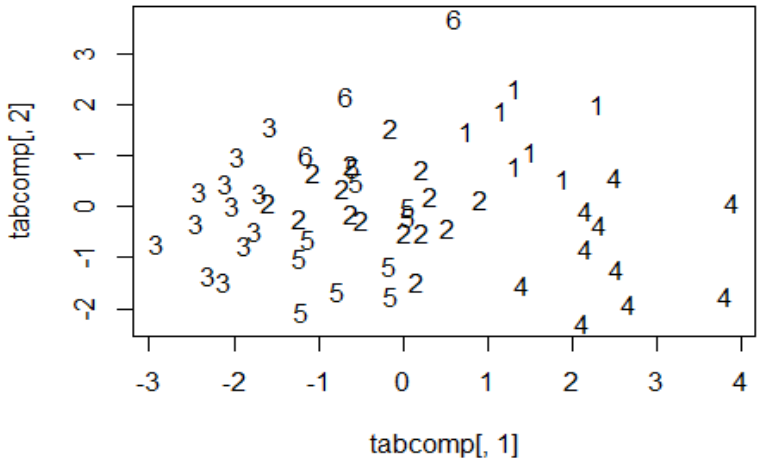
Le coude se situant à l'abscisse ncluster=6, on décide de retenir 6 groupes qu'on peut visualiser sur le dendogramme ci-dessous:

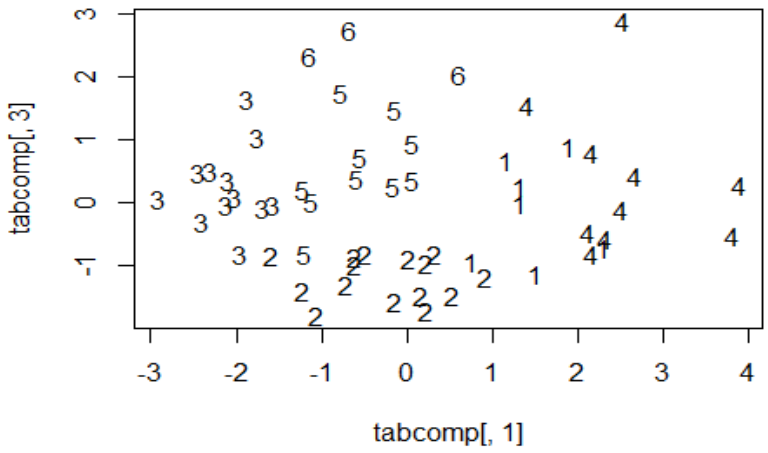


```
#Calcul des R2 par variables
## Dim.1 Dim.2 Dim.3
##  0.88  0.53  0.62


## ANCINS99    MTH70 SHAREFRE VISITSHA CONSUPPA  HONPPAT      AGE
##     0.64     0.58     0.68     0.59     0.37     0.72     0.69
```

On obtient un **R^2 global de 67,43%** pour un CAH sur les 3 composantes de l'ACP alors que le Kmeans de l'exercice précedent nous aurait donné un R2 de 61.9% pour k=6.

```
#Représentation graphique sur les composantes principales (qui contiennent la quasi-totalité de l'information)
```





Groupe 1: axe 1 +, axe 2+, axe 3 =.

Médecin relativement agé avec une bonne expérience ayant un taux de patients de moins de 70 ans relativement faible et se déplaçant relativement peu à domicile.

Groupe 2: axe 1 =, axe 2=, axe 3 –

Médecin d'age et d'expérience moyenne ayant un taux de patients non exonérés de frais médicaux très faible.

Groupe 3: axe 1 –, axe 2=, axe 3 =

Médecin jeune avec peu d'expérience et dont les honoraires sont faibles.

Groupe 4: axe ++, axe 2 -–, axe 3 +.

Médecin très expérimenté et aux honoraires élevés ayant un taux de patients de plus de 70 ans élevé.

Groupe 5: axe 1 =, axe 2 -, axe 3 +

Médecin d'age et d'experience moyenne ayant un taux de patients exonérés de frais médicaux relativement élevé. Taux de patients de plus de 70 ans assez élevé.

Groupe 6: axe -, axe 2 +, axe 3 ++

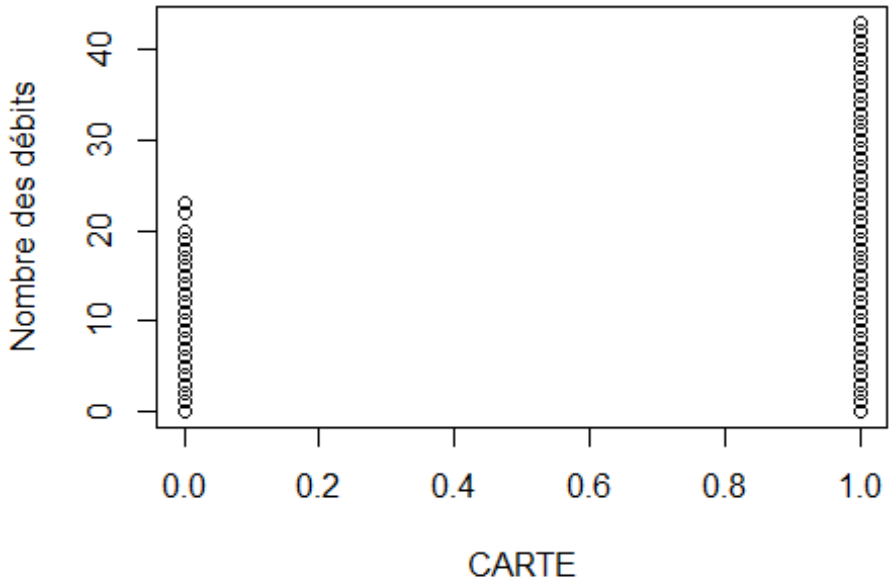
Medecin peu expérimenté et jeune ayant un taux de patients de plus de 70 ans faibble mais un taux de patients exonérés de frais médicaux très elevé.

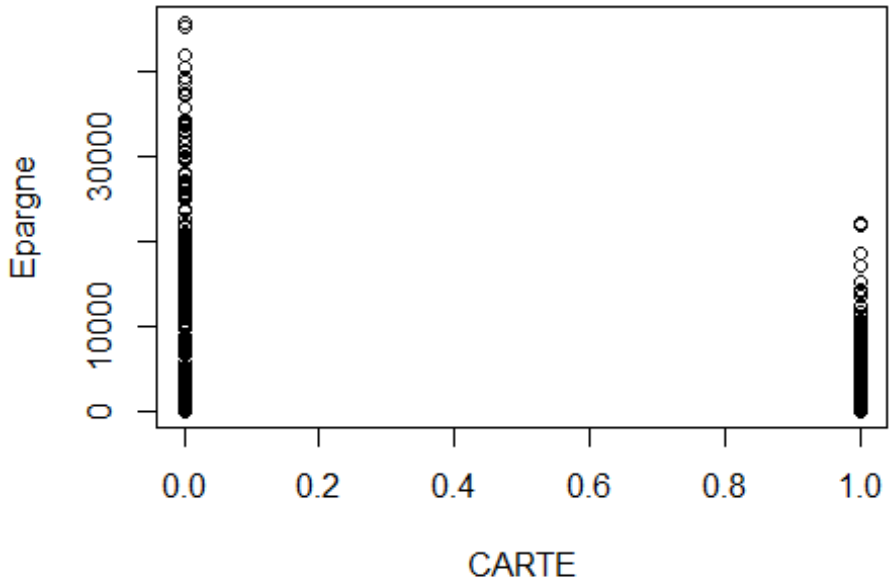
Exercice 2

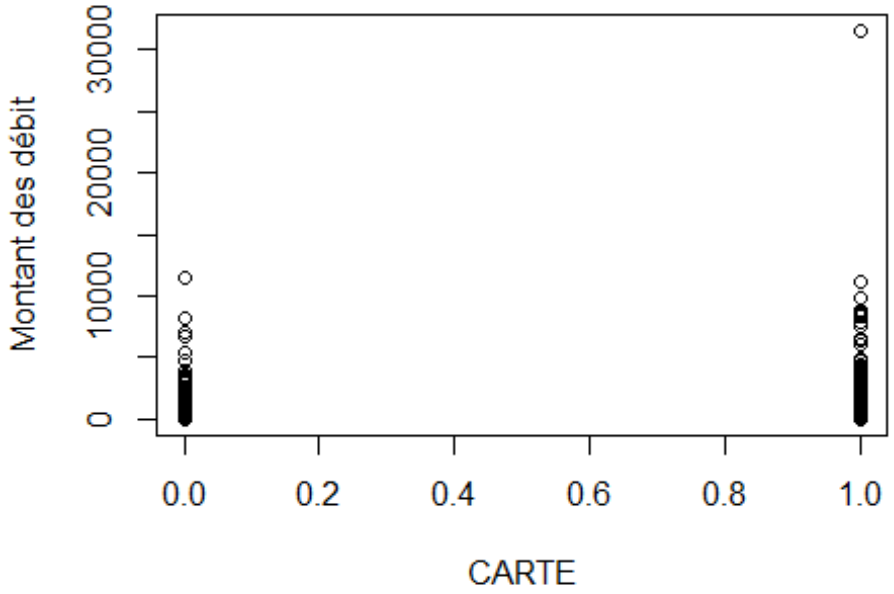
1. Pouvoir discriminants des variables initiales

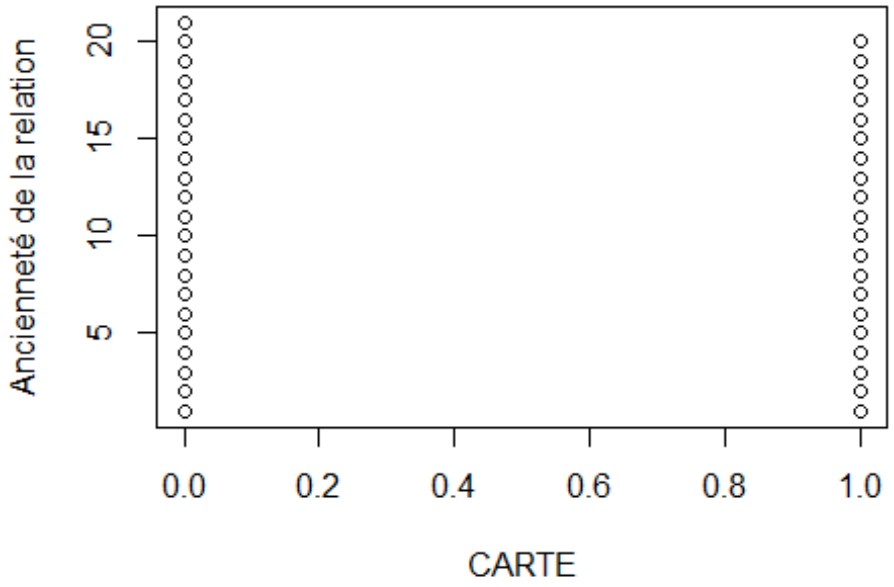
```
## Call:
## lda(CARTE ~ NBRDEB + MTDEB + SOLDCC + ANCREL + EPAR, data = as.data.frame(
scale(banque)))
##
## Prior probabilities of groups:
## -1.00184862274501 0.996925533689748
##         0.4987685         0.5012315
##
```

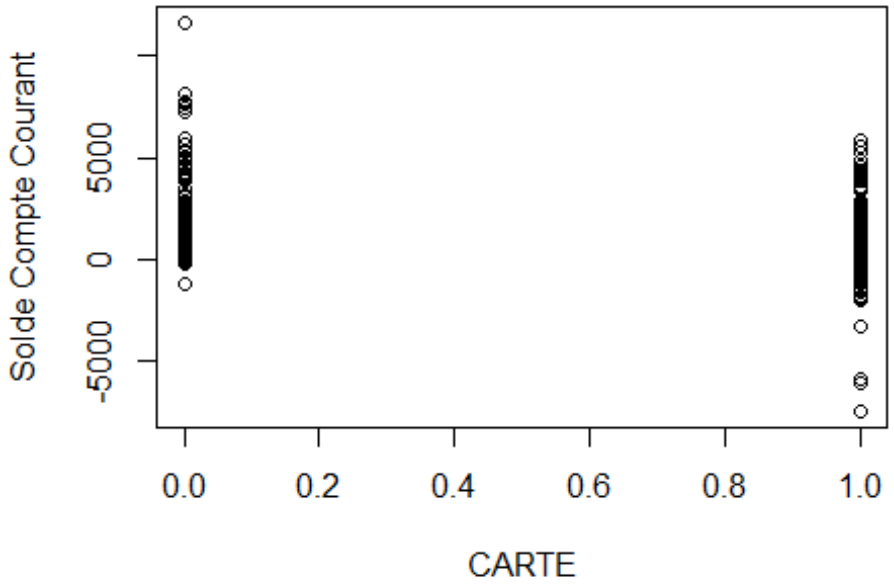
```
## Group means:
##                        NBRDEB      MTDEB      SOLDCC     ANCREL        EPAR
## -1.00184862274501 -0.5088811 -0.2551351  0.02913141  0.1956732  0.4172136
## 0.996925533689748  0.5063804  0.2538813 -0.02898825 -0.1947117 -0.4151634
##
## Coefficients of linear discriminants:
##                 LD1
## NBRDEB  0.98709288
## MTDEB  -0.01788918
## SOLDCC -0.30836366
## ANCREL -0.19259518
## EPAR   -0.59184693
```




Epargne
0
10000
30000
0.0
0.2
0.4
0.6
0.8
1.0
CARTE


Montant des débit
0
10000
20000
30000
0.0
0.2
0.4
0.6
0.8
1.0
CARTE


Ancienneté de la relation
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
CARTE


Solde Compte Courant
-5000
0
5000
0.0
0.2
0.4
0.6
0.8
1.0
CARTE

D après les moyennes des groupes et les graphiques, les variables avec **le plus grand pouvoir discriminant sont NBRDEB et EPAR**. En effet, les personnes possédant la carte Visa Premier semblent avoir peu d'épargne tandis que les personnes ne possédant pas la carte semblent faire moins de retrait. SOLDCC d'après le coefficient semble légèrement discriminant. MTDEB et ANCREL ne semble pas avoir de pouvoir discriminant

2. Echantillonnage

```
index=sample(1:nrow(banque),round(0.8*nrow(banque)))
banquetrain=banque[index,-1]
banquetest=banque[-index,-1]
```

3. AFD sur échantillon d'apprentissage

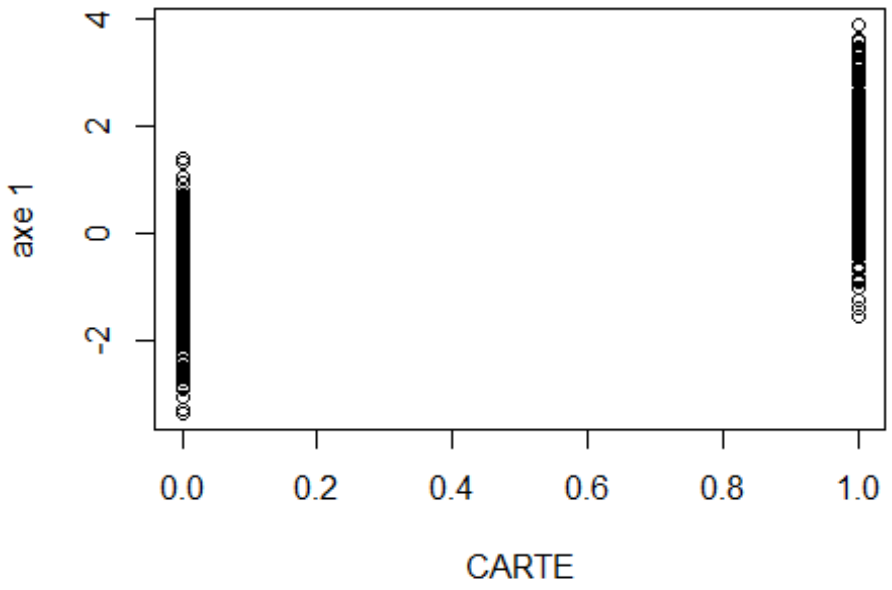
```
## NBRDEB  MTDEB SOLDCC ANCREL   EPAR
##   0.82   0.41  -0.06  -0.32  -0.66
```

Notre axe discriminant est fortement corrélé positivement avec NBRDEB et fortement corrélé négativement avec EPAR. Elle est modérément corrélé positivement avec MTDEB et modérément corrélé négativement avec ANCREL.

```
#Calcul du pouvoir discriminant de notre axe (R2)

## Call:
## lm(formula = axesgroupe$LD1 ~ as.factor(banquetrain$CARTE))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60182 -0.79754  0.05948  0.68887  3.12810
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.76548    0.05547  -13.80   <2e-16 ***
## as.factor(banquetrain$CARTE)1  1.53095    0.07845   19.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 648 degrees of freedom
## Multiple R-squared:  0.3702, Adjusted R-squared:  0.3692
## F-statistic: 380.9 on 1 and 648 DF,  p-value: < 2.2e-16
```

```
# Calcul des moyennes par groupe
Pour les personnes possédant la carte:

## [1] -0.765477

Pour les personnes ne possédant pas la carte:

## [1] 0.765477

#R2=37.16%
```

Notre axe n'a pas un très bon pouvoir discriminant.



Les détenteurs de la carte VISA Premier sont des clients avec un nombre et un montant élevé de mouvement débiteur sur le compte et un solde faible d'épargne. Ils ont une ancienneté de la relation avec la banque plutot faible.

Classement sur l'échantillon d'apprentissage

```
##       1   2
##   0 273  52
##   1 105 220
```

Les clients mal classés sont ceux qui ne sont pas sur la diagonale du tableau ci-dessus. Il y a **161 clients mal classés** et donc le **taux de mal classés est de 24.77%** (161/650)

4. Groupe d'affectation sur échantillon test

5.

```
Classement sur l'échantillon test


##      0  1
##   0 69 11
##   1 25 57
```

Il y a **35 clients mal classés** et donc le **taux de mal classés est de 21.60%** (35/162

ANNEXE: CODE R

#Exercice I

```
#Importation des données
library(FactoMineR)
medecins <- read.delim("C:/Users/chris/Desktop/Stateco/Data Mining/medecins.txt")

#I- KMEANS

#Choix du nombre de groupe

resu_AMM=kmeans(scale(medecins),3,nstart=100) #k=3 R2=39,8%
resu_AMM=kmeans(scale(medecins),4,nstart=100) #k=4 R2=50.0%
resu_AMM=kmeans(scale(medecins),5,nstart=100) #k=5 R2=57.4%
resu_AMM=kmeans(scale(medecins),6,nstart=100) #k=6 R2=61.9%
resu_AMM=kmeans(scale(medecins),7,nstart=100) #k=7 R2=65.5%
resu_AMM=kmeans(scale(medecins),8,nstart=100) #k=8 R2=68.9%

vectgroupe=c(3,4,5,6,7,8)
vectR2=c(39.8,50,57.4,61.9,65.5,68.9)
plot(vectgroupe,vectR2)
lines(vectgroupe,vectR2)

resuAMM=kmeans(scale(medecins),5,nstart=100)
groupe=as.factor(resuAMM$cluster)
medecins2=cbind(medecins,groupe)
```

```
#Calcul des R2 par variable
tabR2=catdes(medecins2,8,proba=1)$quanti.var#
tabR2

medecins3=medecins2[,c(-2,-5)]

moyech=round(apply(medecins3[,-6],2,mean),digits=2)
moygr1=round(apply(medecins3[groupe==1,-6],2,mean),digits=2)
moygr2=round(apply(medecins3[groupe==2,-6],2,mean),digits=2)
moygr3=round(apply(medecins3[groupe==3,-6],2,mean),digits=2)
moygr4=round(apply(medecins3[groupe==4,-6],2,mean),digits=2)
moygr5=round(apply(medecins3[groupe==5,-6],2,mean),digits=2)
tabmoy=rbind(moygr1,moygr2,moygr3,moygr4,moygr5,moyech)

tabmoy

#Graphique représentant les groupes

plot(medecins3$AGE,medecins3$HONPPAT,type='n')
text(medecins3$AGE,medecins3$HONPPAT,labels=medecins3$groupe)
```

#II- CAH sur CP

```
#ACP
resuacp=PCA(medecins)
resuacp$eig
plot(resuacp$eig[,1])
```

```
lines(resuacp$eig[,1])
tabcomp=resuacp$ind$coord[,1:3]

#CAH sur les composantes principales retenues (inutile de standardiser)
resuCAHmedecins= hclust(dist(tabcomp),method="ward.D2")
plot(resuCAHmedecins)
ncluster=56:1
plot(ncluster,resuCAHmedecins$height)
lines(ncluster,resuCAHmedecins$height)

#Dendrogramme

plot(resuCAHmedecins)
rect.hclust(resuCAHmedecins,k=6)

groupecahmedecins=cutree(resuCAHmedecins,k=6)

#fonction de calcul R2

calculR2=function(variable,groupe)
{
 groupe=as.factor(groupe)
 resuanova=anova(lm(variable~groupe))
 scinter=resuanova$Sum[1]
 sct=sum(resuanova$Sum)
 R2=scinter/sct
 R2
}
```

```
#Calcul de la scinter et de la sct

calculsc=function(variable,groupe)
{
 variable=scale(variable)
 groupe=as.factor(groupe)
 resuanova=anova(lm(variable~groupe))
 scinter=resuanova$Sum[1]
 sct=sum(resuanova$Sum)
 resu=c(scinter,sct)
 resu
}

#Calcul des R2 par variables
round(apply(tabcomp,2,calculR2,groupecahmedecins),digits=2)

round(apply(medecins,2,calculR2,groupecahmedecins),digits=2)

tabsc=apply(tabcomp,2,calculsc,groupecahmedecins)
somme=apply(tabsc,1,sum)
inertieinter=somme[1]
inertie=somme[2]
R2global=inertieinter/inertie

#Représentation graphique sur les composantes principales (qui contiennent la quasi-totalité de l'information)
```

```
plot(tabcomp[,1],tabcomp[,2],type='n')
text(tabcomp[,1],tabcomp[,2],labels=groupecahmedecins)

plot(tabcomp[,1],tabcomp[,3],type='n')
text(tabcomp[,1],tabcomp[,3],labels=groupecahmedecins)
```

#Exercice 2

```
#importation des données
library(MASS)
banque <- read.csv("C:/Users/chris/Desktop/Stateco/Data Mining/banque.txt", sep="")

resu_afd=lda(CARTE~NBRDEB+MTDEB+SOLDCC+ANCREL+EPAR,data=banque)

plot(banque$CARTE,banque$NBRDEB,xlab="CARTE",ylab="Nombre des débits")
plot(banque$CARTE,banque$EPAR,xlab="CARTE",ylab="Epargne")
plot(banque$CARTE,banque$MTDEB,xlab="CARTE",ylab="Montant des débit")
plot(banque$CARTE,banque$ANCREL,xlab="CARTE",ylab="Ancienneté de la relation")
plot(banque$CARTE,banque$SOLDCC,xlab="CARTE",ylab="Solde Compte Courant")

index=sample(1:nrow(banque),round(0.8*nrow(banque)))
banquetrain=banque[index,-1]
banquetest=banque[-index,-1]

resu_afd2=lda(CARTE~.,data=banquetrain)

axes=predict(resu_afd2)$x
```

```
#Matrice de correlation

banque2train=banquetrain[,-4] #enleve la variable CLIENT et CARTE
mat=cbind(banque2train,axes) #rajoute notre axe discriminant
matcor=cor(mat)
matcor=round(matcor,digits=2)
matcor=matcor[-c(6),c(6)]
matcor

#Calcul du pouvoir discriminant de notre axe (R2)

axesgroupe=cbind(axes,banquetrain)
summary(lm(axesgroupe$LD1~as.factor(banquetrain$CARTE)))
#R2=37.16%

#Calcul de la moyenne par groupe
tableau=cbind(axesgroupe$LD1,axesgroupe$CARTE)
mean(tableau[axesgroupe$CARTE==0,1])
mean(tableau[axesgroupe$CARTE==1,1])

#Graphique des groups

plot(axesgroupe$CARTE,axesgroupe$LD1,,xlab="CARTE",ylab="axe 1")
```

```
#Validation de l'AFD

probas=predict(resu_afd2)$posterior
classement= apply(probas,1,which.max)
table(banquetrain$CARTE,classement)

probatest=predict(resu_afd2,newdata=banquetest)
table(banquetest$CARTE,probatest$class)
```