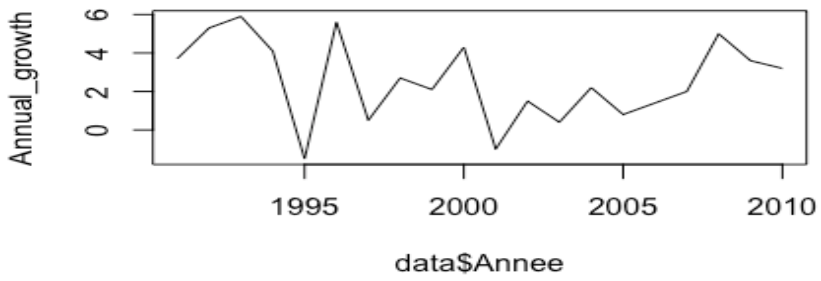
Gregoire Lecallier 31001484

DEVOIR MAISON ANALYSE DE DONNESS

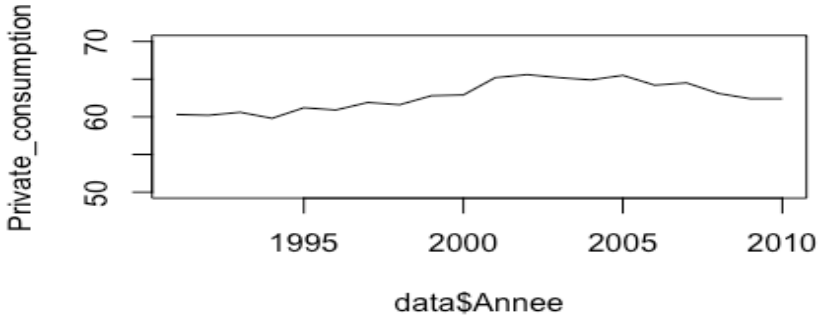
1. Analyse Descriptive :

Effectuons une analyse descriptive des données entre 1991 et 2010.

En ce qui concerne le taux de croissance dans cette période, la moyenne est de 2.59%. Cependant il plus pertinent de calculer pour cette variable le taux de croissance annuel moyen c'est à dire le taux de croissance qu'il aurait fallu avoir chaque année depuis 1991 pour arriver au niveau de PIB de 2010. Il est de 2.57%.
Ci-dessous la représentation graphique de la variable croissance :

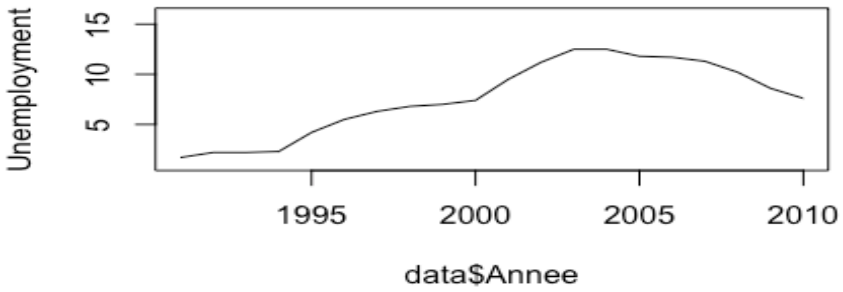


Le taux de consommation moyen privé est de 62.76 c'est à dire qu'en moyenne la population consacre au Royaume Unis 63% de son revenu à la consommation, le reste (37%) probablement consacré à l'épargne. On remarque qu'il n'y a pas beaucoup de différence entre le min (60) et le max (65.6) ce qui laisse penser que le taux est relativement constant au cours du temps.
En effet la variance est de 3.7 ce qui est peu pour cette période de temps, la part allouée à la consommation chaque année par les ménages est donc globalement stable entre 1991 et 2010 comme on peut le voir dans le graphique ci-dessous :

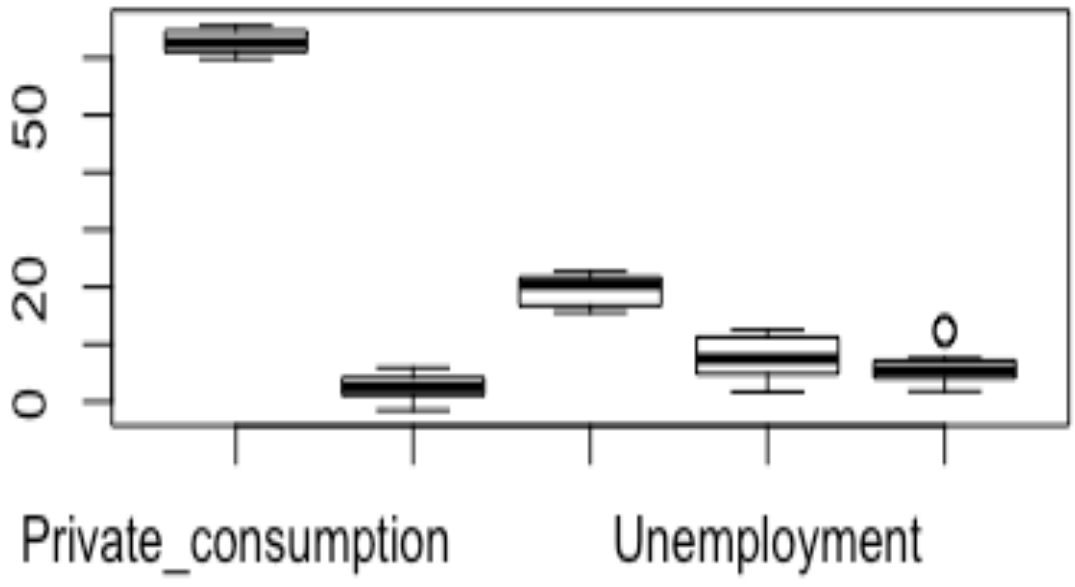


Gregoire Lecallier 31001484

Le taux de chômage moyen est de 7.6% pour la période 1991-2010. Il y a un grand écart entre le taux de chômage minimum (1.7%) et le taux de chômage maximum (12.5%) ce qui laisse penser une grande variabilité du taux de chômage.
La variance est élevée (13.7) ce qui confirme une forte variabilité du chômage au RU entre 1991 et 2010.



L'inflation moyenne est de 5.8% pour la période 1991-2010. Sa variabilité est plutôt élevée si l'on prend en considération les valeurs « aberrantes » qui constituent un pic autour des 12% au milieu des années 90.
SI l'on retire ces 2 valeurs, la faible variabilité de l'inflation est confirmée par le boxplot écrasé ci-dessous :

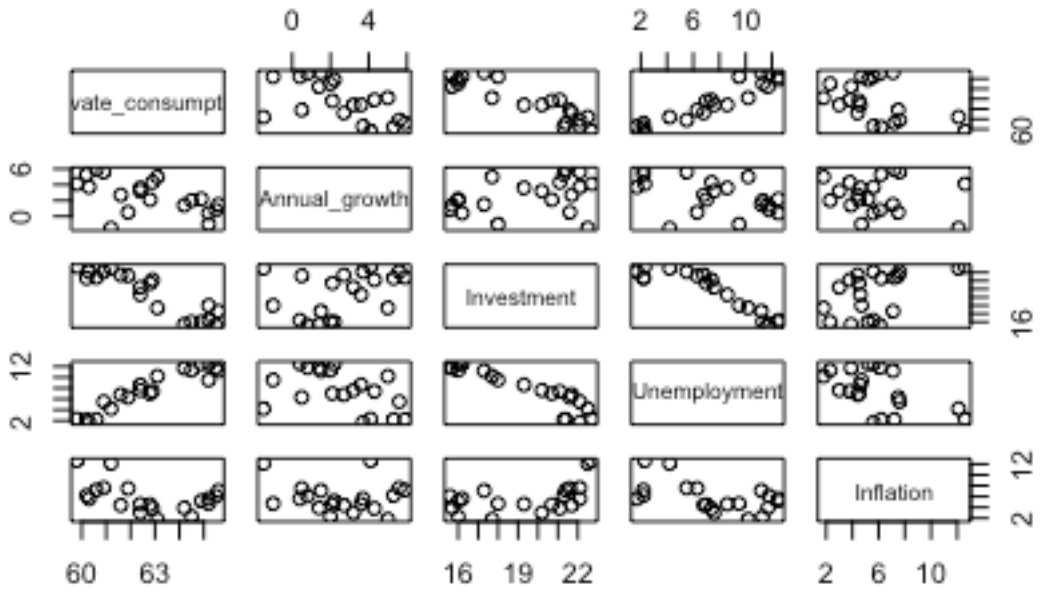


Gregoire Lecallier 31001484

Le boxplot ou boite à moustache, qui représente la variabilité des valeurs, confirme la stabilité des valeurs prises par le taux de consommation et la dispersion des valeurs prises par le taux de chômage.
Les variables qui ont le moins de variabilité sont donc le taux de consommation et la croissance annuelle.

Le coefficient de corrélation est positif et élevé entre le taux de chômage et la consommation privée (0.94), les deux variables évoluent donc dans le même sens et semblent être très liées de façon linéaire.
A l'inverse les variables Investissement et taux de consommation sont très liées négativement (-0.9), elles évoluent en sens opposé. Il en est de même pour le lien entre chômage et investissement (-0.9).
On observe une indépendance linéaire entre Inflation et taux de croissance (coefficient de corélation linéaire faible = -0.1).

On représente les liens entre les variables par le graphique de corrélation suivant :



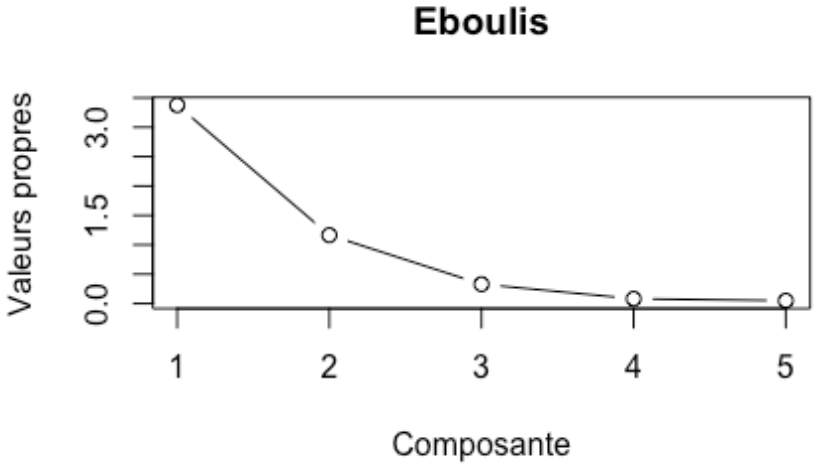
Gregoire Lecallier 31001484

2. Analyse en composante principale :

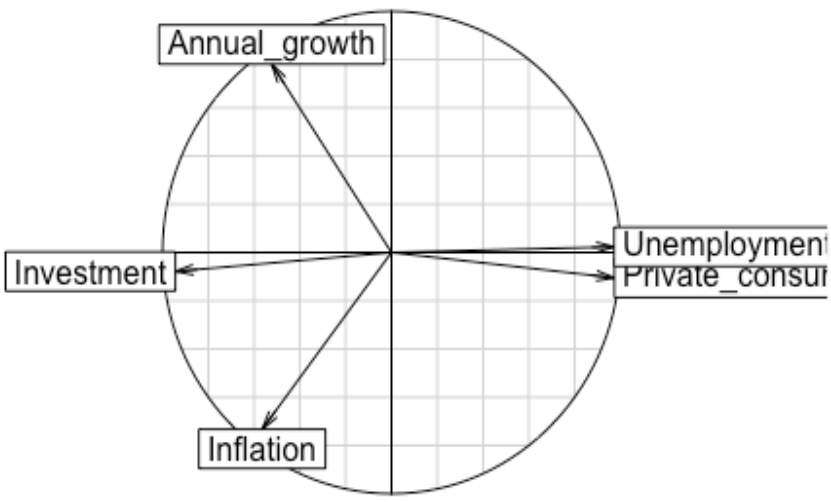
a)Le jeu de données contient 5 variables, nous souhaitons réduire le nombre de données en les regroupant et créer ainsi de nouvelles variables. Avec une ACP nous allons ainsi expliciter les liaisons entre les variables et décrire les données avec des nouvelles variables.
Le but est donc d'expliquer les anciennes variables grâce aux nouveaux axes crées à l'issue de l'ACP.

b) On choisit d'effectuer une ACP centrée réduite car bien que les variables soient toutes des taux, elles n'ont pas le même ordre de grandeur, la consommation et l'investissement étant de l'ordre des dizaines et le reste de l'ordre de l'unité.

Nous pouvons garder les 2 premiers axes qui expliquent 90% de la variabilité d'après les proportions de variance « captées »par ces 2 axes.
En effet les valeurs propres des 2 premiers axes sont supérieures à 1, en ACP normée on garde en générale celles-ci.
On représente ci-dessous les éboulis des axes qui représente l'information captée par chacun des 5 axes de l'ACP : le premier axe capte à lui seul la majeure partie de l'information.



c)Cercle de corrélation :



Gregoire Lecallier 31001484

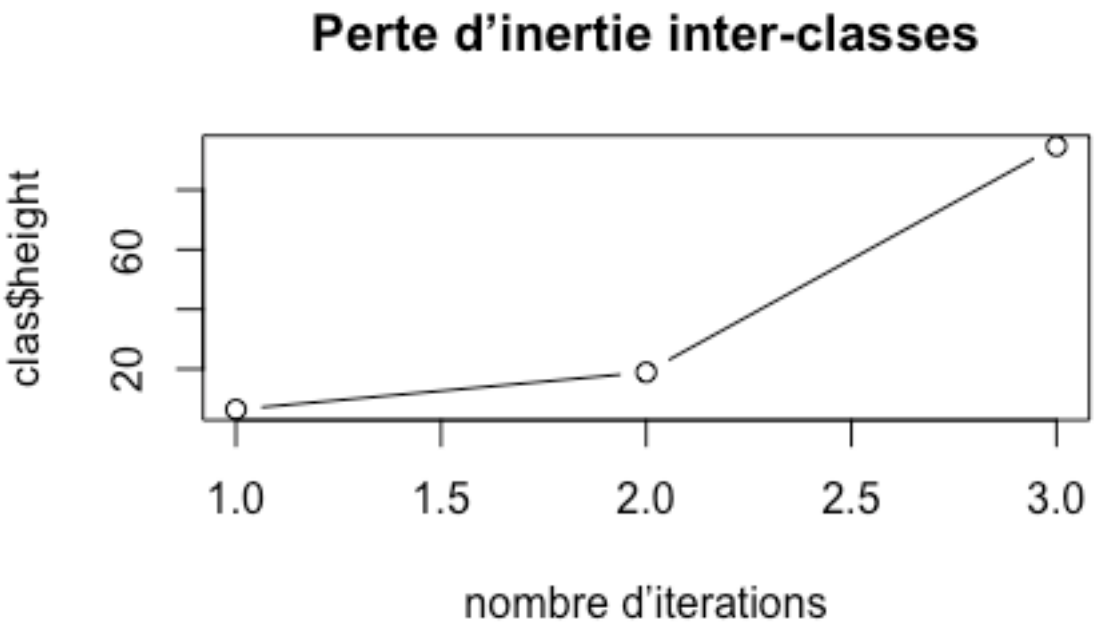
Le cercle de corrélation représente en abscisse, la corrélation entre le nouvel axe 1 et la variable de la table de départ et en ordonnée, la corrélation entre l'axe 2 et la variable de la table.

On remarque que l'axe 1 est presque confondu avec le taux de chômage, la corrélation entre ces deux derniers est très proche de 1. L'axe 1 représente donc en autre le taux de chômage. Il en est de même dans une moindre mesure pour le taux de consommation et l'investissement (corrélation négative pour l'investissement).
L'axe 2 quant à lui est plus corrélé à l'inflation et au taux de croissance.

d)Plus les observations sont situées à gauche de l'axe 1 plus elles auront un taux d'investissement faible (car corrélation entre l'axe 1 et investissement proche de -1).
Les "individus" à droite de l'axe 1 auront un taux de chômage et une croissance élevée.
Notons l'observation d'un effet taille sur ces 2 dernieres variables (variables très corrélées).
L'axe 1 distingue donc les variables avec un faible et fort taux d'investissement, de chômage et de consommation.
L'axe 2 distingue de la même manière les variables avec un faible et fort taux de croissance et d'inflation.

3. <u>Classification hiérarchique des données :</u>

a)La distance utilisée est la distance entre individus, c'est la distance euclidienne.
La distance de ward est la distance entre les centres des deux groupes que l'on compare.
La méthode de WARD fait la distance entre les centres de classes
L'information est représentée par l'inertie interclasses :



```
data = read.csv("/Users/gregoirelecallier/Desktop/UPX/ADD/data_Roy-
 Uni.csv",header = TRUE, sep =";")
data1 = data[2:6]
attach(data1)
#Question 1:

summary(data1)
#entre 1991 et 2010
#Taux de croissance :
plot(Annual_growth~data$Annee,type = "l")
n = 20
TCAM = (prod((Annual_growth/100)+1)^(1/n)-1)*100 #C'est le taux de
 croissance qu'il aurait fallu chaque année depuis 1991 pour arriver
 au niveau de PIB de 2010
TCAM
var(Annual_growth)

#Consommation privée :
plot(Private_consumption~data$Annee,type = "l",ylim=c(50,70))
var(Private_consumption)

#Chomage:
var(Unemployment) #Variance élevée ce qui confirme une forte
 variabilité du chômage au RU entre 1991 et 2010.
plot(Unemployment~data$Annee,type = "l",ylim = c(1,16))

#Inflation:
var(Inflation)
#On retire les valeurs abérantes :
Inflation1 = Inflation[-4]
Inflation2 = Inflation1[-4]
var(Inflation2)
plot(Inflation~data$Annee,type = "l")

boxplot(data1)
cor(data1)
plot(Unemployment~Private_consumption,col=grey(0.6))
plot(Investment~Private_consumption,col=grey(0.6))
plot(Inflation~Private_consumption,col=grey(0.6))
pairs(data1) #Représentation générale des liaisons entre les
 variables

#Question 2 :
#b)
boxplot(data1)
res.acp = prcomp(data1, scale=TRUE)
summary(res.acp)
```

```
val.propres = res.acp$sdev^2
plot(val.propres,type = "b",ylab="Valeurs
 propres",xlab="Composante",main="Eboulis") #Eboulis des valeurs
 propres

z = res.acp$x
boxplot(z)
#c)d)
cc = cor(data1,z[,1:2])
cc
s.corcircle(cc,lab=names(data1))

#question 3:

tdata= data.frame(t(data1))
names(tdata) = tdata[1,]
tdata1=tdata[-1,]
tdata=tdata1
# perte d'inertie :
d=dist(scale(tdata)) #sur les donnees centrees reduites
clas=hclust(d^2,meth="ward")
plot(clas$height,type="b",main="Perte d'inertie inter-classes",
 xlab="nombre d'iterations")
#et le dendogramme :
clas=hclust(d^2,meth="ward")
plot(clas)
#etape 0 50 groupes
#etape 49 1 groupe
#trouver le juste milieu
##pour trois groupes on a la variabilité entre les classes
gr3 = cutree(clas,k=3)
table(gr3)
```