Projet analyse de donnee sur le cinema

mohamed adam bacouch

2023-05-01

Introduction:

Dans ce projet, nous effectuerons une analyse de données de notre base de données “cinéma”. Cet ensemble contient environ 130 observations, avec différents sexes, catégories socioprofessionnelles et bien d’autres. À la fin, nous tenterons d’étudier les différentes raisons pour lesquelles les gens vont au cinéma et s’il existe une certaine corrélation entre la raison d’allers au cinema et les autres variables de notre ensemble de données.

Appel des bibliotheques :

```
suppressPackageStartupMessages({
library(dplyr)
library(lessR)
library(FactoMineR)
library(factoextra)
library(corrplot)
library(RColorBrewer)
library(readxl)
})
```

```
## Warning: le package 'lessR' a été compilé avec la version R 4.2.3
```

Importation de la base de donnée:

```
cinema<- read_excel("C:/Users/poste/Desktop/new_cinema.xlsx")
head(cinema)
```

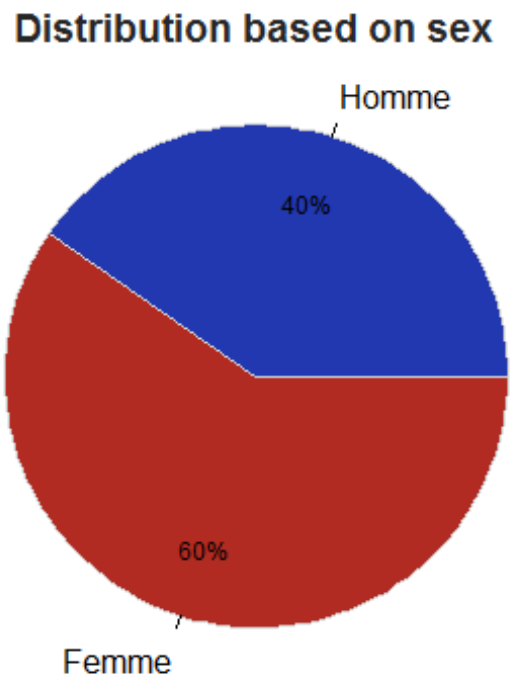
```
## # A tibble: 6 x 71
##   Sexe  Catégori~1 age   John ~2 the d~3 titanic ce qu~4 the p~5 annab~6
ratat~7
##   <chr> <chr>      <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1 Homme Etudiant   Entr~       4       3       3       4       5       3
3
## 2 Homme Etudiant   Entr~       5       4       3       1       2       3
3
## 3 Homme Etudiant   Entr~       1       4       4       1       1       1
1
## 4 Homme Etudiant   Entr~       1       1       5       1       1       1
1
```

```
## 5 Femme Etudiant   Entr~       1       3       5       5       1       3
4
## 6 Homme Etudiant   Entr~       5       5       5       4       5       2
4
## # ... with 61 more variables: `12 years a slave` <dbl>, interstellar
<dbl>,
## #   `harry potter` <dbl>, `the matrix` <dbl>, `home alone` <dbl>,
## #   `the notebook` <dbl>, gladiator <dbl>, braveheart <dbl>,
## #   `the conjuring` <dbl>, `kung fu panda` <dbl>, `the greenmile` <dbl>,
## #   inception <dbl>, `lord of the rings` <dbl>, `the dark knight` <dbl>,
## #   `the hangover` <dbl>, `the fault in our stars` <dbl>, troy <dbl>,
## #   saw <dbl>, shrek <dbl>, `the godfather` <dbl>, ...
```

1. statistiques descriptives

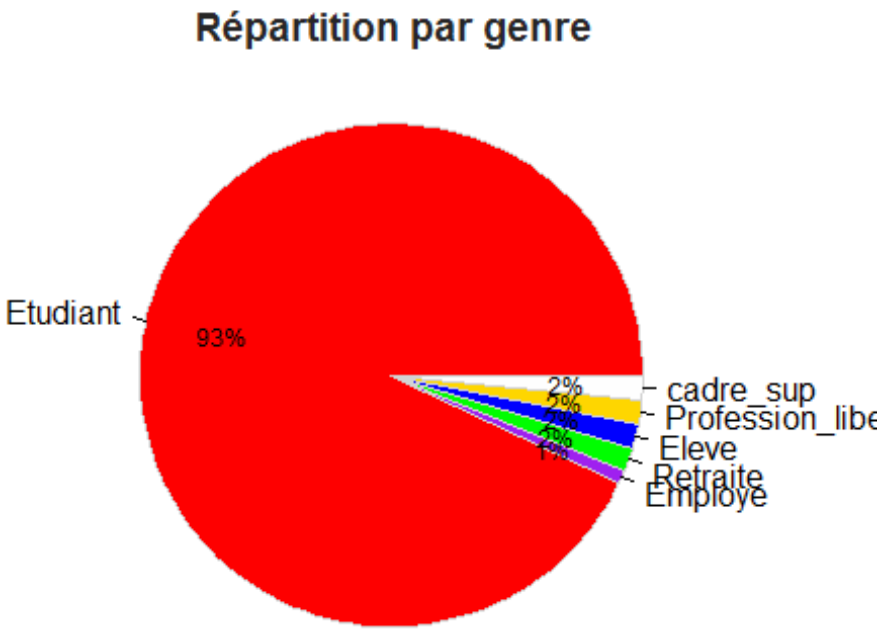
1.1 Diagramme circulaire de la distribution selon le sexe

```
Homme=sum(cinema$Sexe == "Homme")
Femme=sum(cinema$Sexe =="Femme")

n=c(Homme,Femme)
Sexe=as.factor(c("Homme","Femme"))

gender=as.data.frame(cbind(Sexe,n))
PieChart(x=Sexe,y=n,gender,fill =c("#2238B2","#B22B22"),
         values_color="Black",hole=0,main="Distribution based on sex")
```



1.2 Diagramme circulaire de la distribution selon catégorie socioprofessionnelle

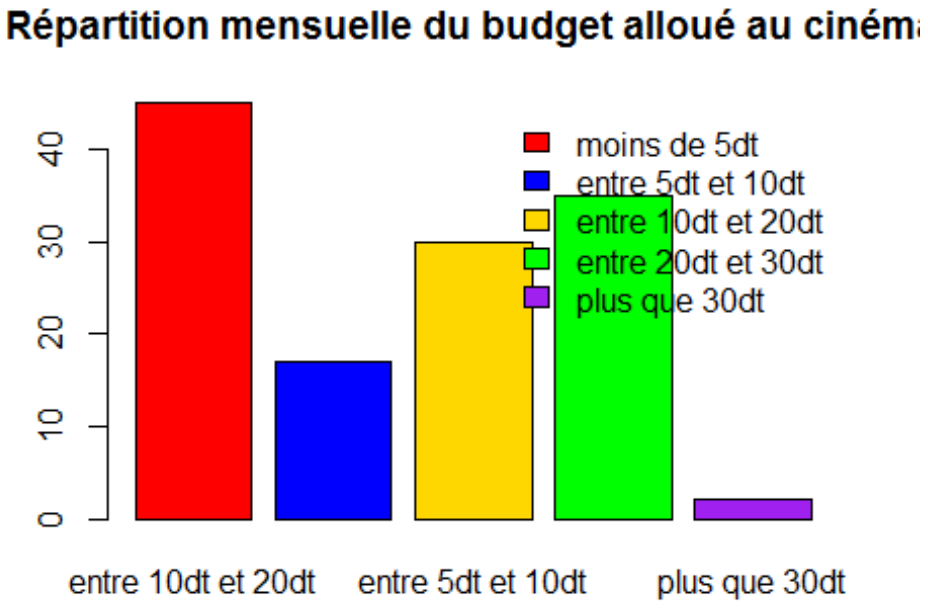
```
Etudiant=sum(cinema$Catégorie_socioprofessionnelle == "Etudiant")
Employe=sum(cinema$Catégorie_socioprofessionnelle =="Employé")
Retraite=sum(cinema$Catégorie_socioprofessionnelle =="Retraité")
Eleve=sum(cinema$Catégorie_socioprofessionnelle =="Elève")
Profession_liberale=sum(cinema$Catégorie_socioprofessionnelle =="Profession
libérale")
cadre_sup=sum(cinema$Catégorie_socioprofessionnelle =="Cadre et profession
intellectuelle supérieure")
n=c(Etudiant,Employe,Retraite,Eleve,Profession_liberale,cadre_sup)
prof=as.factor(c("Etudiant","Employe","Retraite","Eleve","Profession_liberale
","cadre_sup"))

Job=as.data.frame(cbind(prof,n))
PieChart(x=prof,y=n,Job,fill =c("red", "purple", "green", "blue",
"gold","white"),
         values_color="Black",hole=0,main=" Répartition par genre")
```



1.3 Répartition mensuelle du budget alloué au cinéma

```
y=table(cinema[,69])
ll=c("moins de 5dt","entre 5dt et 10dt","entre 10dt et 20dt","entre 20dt et
30dt","plus que 30dt")
barplot(y,col=c("red","blue","gold","green","purple"),main="Répartition
mensuelle du budget alloué au cinéma")
```

```
legend(x="topright",legend=ll,fill=c("red","blue","gold","green","purple"),bty="n")
```



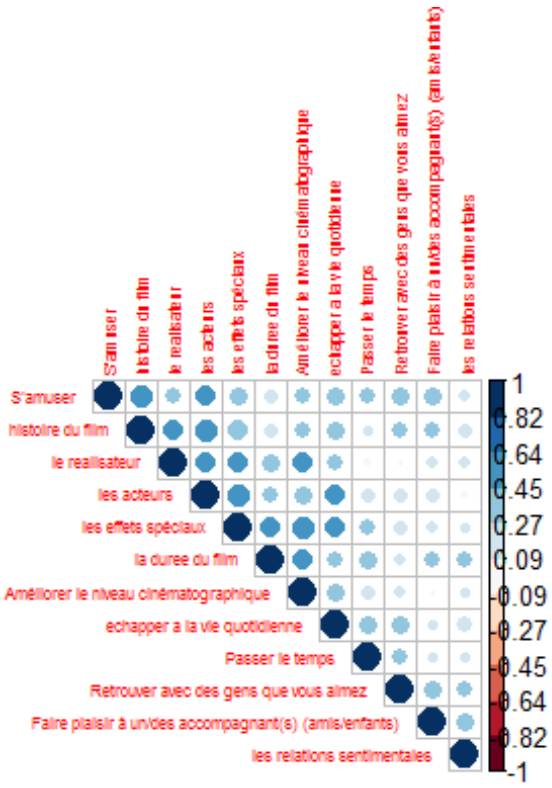
2.Analyse en composantes principales des raisons d'aller au cinéma:

Dans cette partie On va effectuer nos études sur les colonnes qui contiennent des informations concernant les raisons d'aller au cinéma:

```
reason=as.matrix(cinema[,c(57:68)])
```

2.1 Pertinence de l'ACP:

```
R<-cor(reason)
corrplot(R, type = "upper", order = "hclust", col = brewer.pal(n = 11, name = "RdBu"),
         tl.cex = 0.5)
```

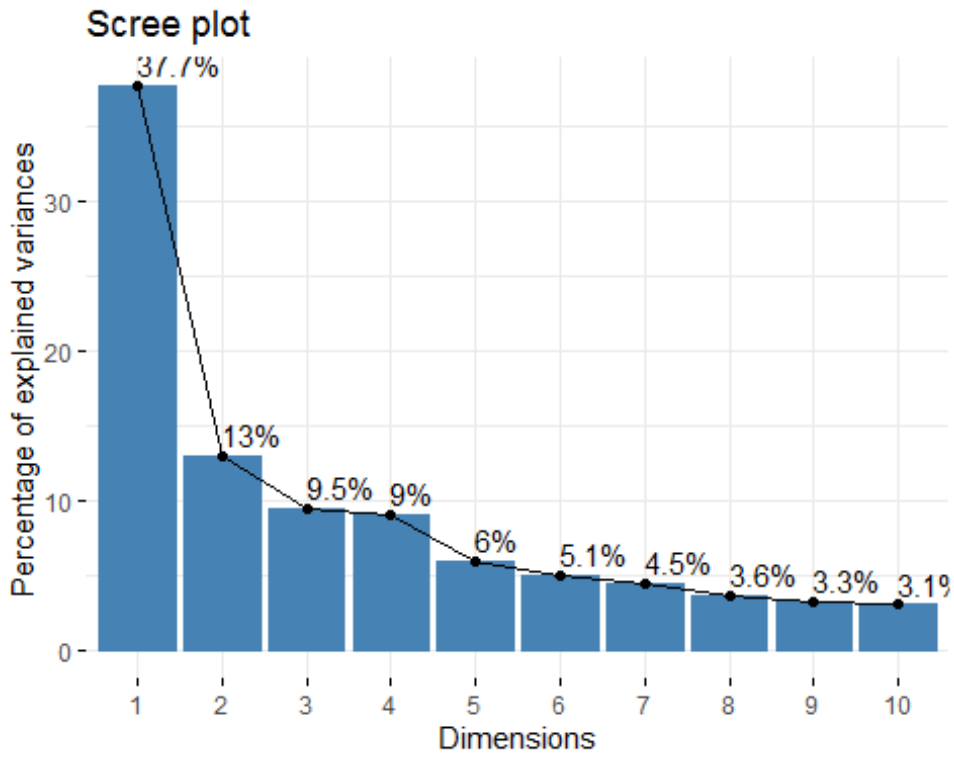


2.2 Choix de nombre d'axes:

```
res.pca_R=PCA(reason,ncp = 4,graph= F)
head(res.pca_R$eig)
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1  4.5284349              37.736957                           37.73696
## comp 2  1.5553176              12.960980                           50.69794
## comp 3  1.1402565               9.502137                           60.20007
## comp 4  1.0851903               9.043253                           69.24333
## comp 5  0.7215584               6.012986                           75.25631
## comp 6  0.6069428               5.057857                           80.31417
```

```
fviz_eig(res.pca_R ,addlabels = TRUE)
```



Interpretation:

-Critère de Kaiser : En observant les valeurs propres des axes, nous pouvons constater que 4 d'entre eux ont des valeurs supérieures à 1. Par conséquent, nous retenons ces 4 axes en accord avec ce critère.

-Critère du taux d'inertie cumulée : En observant le taux d'inertie cumulé des 2 premiers axes, qui atteint environ 50%, nous constatons que celui-ci est significatif. Nous décidons donc de retenir uniquement ces 2 axes pour notre analyse

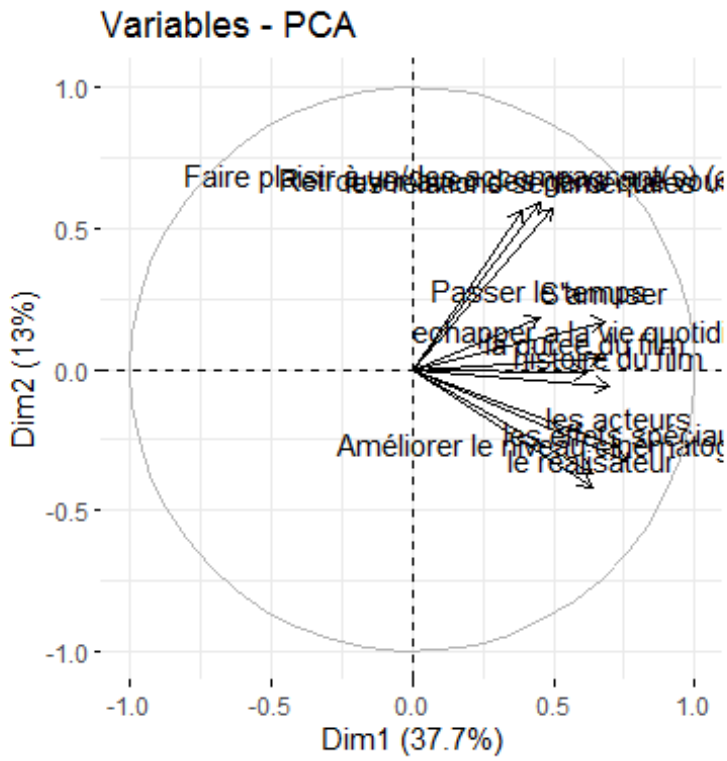
-Critère du coude : En observant le graphique, on peut remarquer que le point de coude est situé au niveau du deuxième axe. Par conséquent, nous choisissons de retenir deux axes pour notre analyse.

=>En combinant les 3 critères, il serait judicieux de retenir les 2 premiers axes.

2.3 Interpretation de la carte des variables :

```
res.pca_R$var$coord[,1:2]

##                                                          Dim.1
Dim.2
## S'amuser                                              0.6808707
0.172232791
## Faire plaisir à un/des accompagnant(s) (amis/enfants) 0.4501823
0.591827989
## Retrouver avec des gens que vous aimez                0.4983747
0.571462722
```

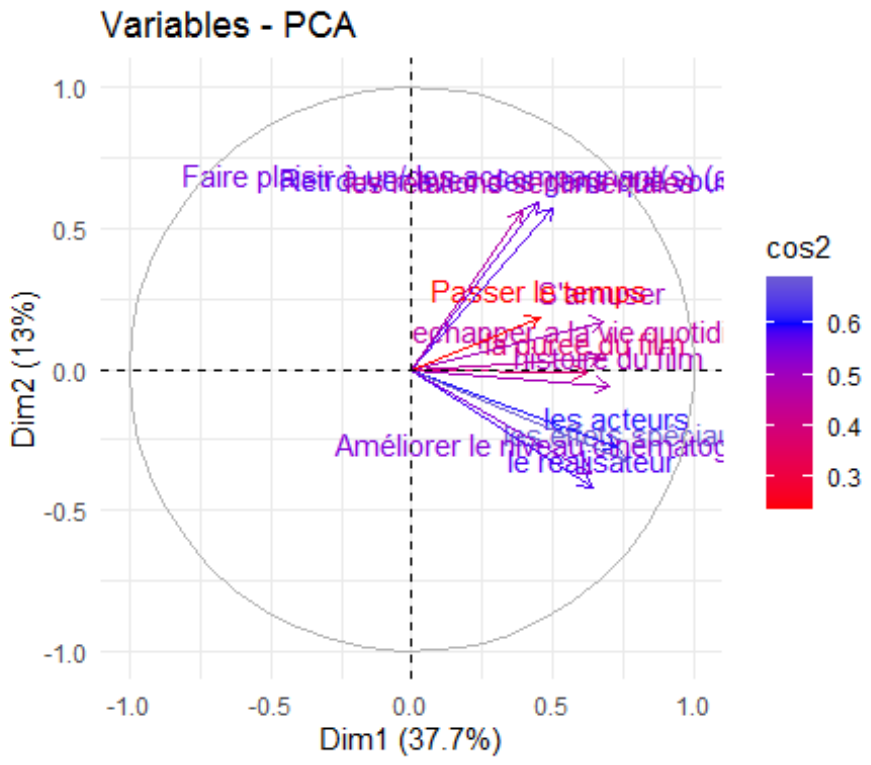
```
## les relations sentimentales                                0.3898834
0.564366287
## echapper a la vie quotidienne                              0.6761586
0.039336848
## histoire du film                                           0.6985886 -
0.057691565
## les acteurs                                                0.7296063 -
0.274109547
## les effets spéciaux                                        0.7638172 -
0.327605224
## le realisateur                                             0.6380044 -
0.418216999
## la duree du film                                           0.6189467 -
0.009742837
## Améliorer le niveau cinématographique                      0.6375101 -
0.367162911
## Passer le temps                                            0.4519705
0.182121348

fviz_pca_var(res.pca_R,shadow=TRUE)
```




```
res.pca_R$var$cos2[,1:2]

##                                                              Dim.1
Dim.2
## S'amuser                                                   0.4635849
0.02966413422
```

```
## Faire plaisir à un/des accompagnant(s) (amis/enfants) 0.2026641
0.35026036894
## Retrouver avec des gens que vous aimez                0.2483774
0.32656964268
## les relations sentimentales                           0.1520091
0.31850930595
## echapper a la vie quotidienne                         0.4571904
0.00154738763
## histoire du film                                      0.4880260
0.00332831671
## les acteurs                                           0.5323253
0.07513604370
## les effets spéciaux                                   0.5834167
0.10732518308
## le realisateur                                        0.4070496
0.17490545848
## la duree du film                                      0.3830950
0.00009492288
## Améliorer le niveau cinématographique                 0.4064191
0.13480860333
## Passer le temps                                       0.2042773
0.03316818544
```

```
fviz_pca_var(res.pca_R, col.var = "cos2")+
  scale_color_gradient2(low="red" ,mid="blue" ,
                        high="green", midpoint = 0.6)+
  theme_minimal()
```



Interpretation :

-Les variables “les acteurs”, “Améliorer le niveau cinématographique”, “Le réalisateur” et “Les effets spéciaux” sont corrélées entre elles et également corrélées avec l’axe 1. Nous pouvons regrouper ces variables en une seule variable appelée “Éléments de production cinématographique”.

-Les variables “L’histoire du film”, “La durée du film”, “Échapper à la vie quotidienne”, “Passer le temps” et “S’amuser” sont corrélées entre elles et également Fortement corrélées avec l’axe 1. Nous pouvons regrouper ces variables en une seule variable appelée “Plaisir de film”.

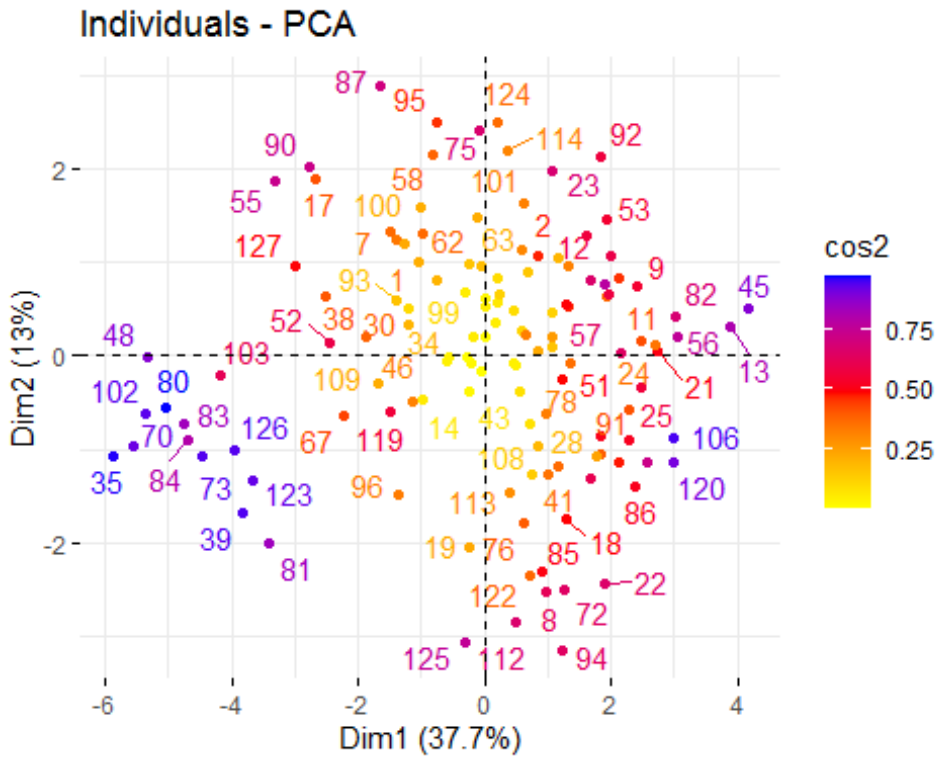
-Les variables “Faire plaisir à un/des accompagnant(s) (amis/enfants)”, “Retrouver avec des gens que vous aimez”, “Les relations sentimentales” sont corrélées entre elles et également corrélées avec l’axe 2. Nous pouvons regrouper ces variables en une seule variable appelée “Plaisir Social”.

-l’axe 1 favorise “Éléments de production cinématographique” et “Plaisir de film”, il est donc légitime de nommer l’axe 1 : “Expérience cinématographique”.

-l’axe favorise “Plaisir Social”, il est donc légitime de nommer l’axe 2 : “Expérience Social”.

Carte des Individus:

```
fviz_pca_ind(res.pca_R, col.ind = "cos2",
             gradient.cols = c("yellow", "red", "blue"),
             repel = TRUE
             )
```

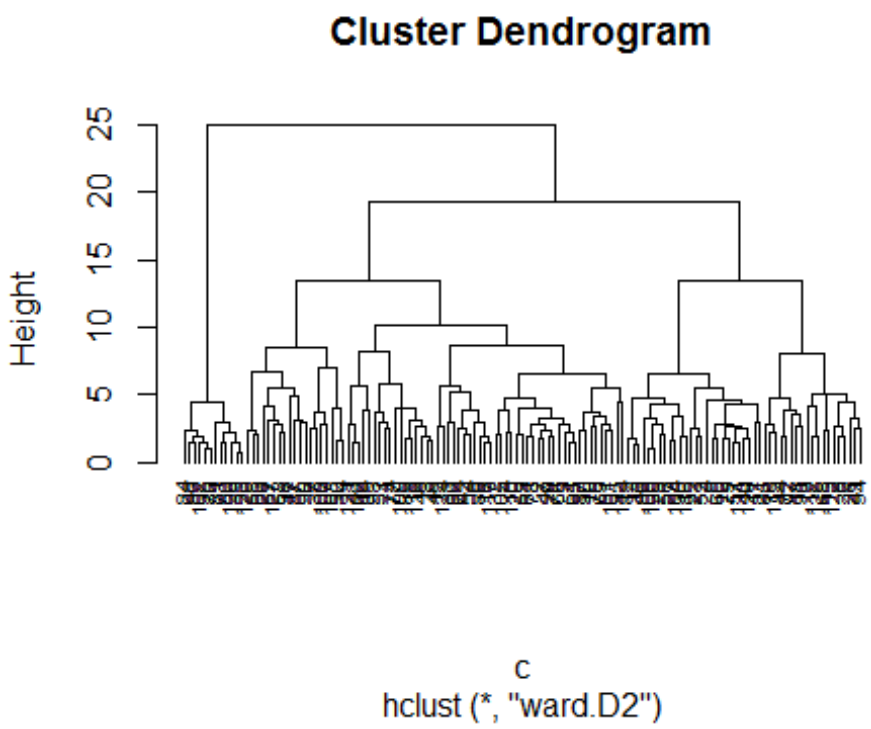


Interpretation :

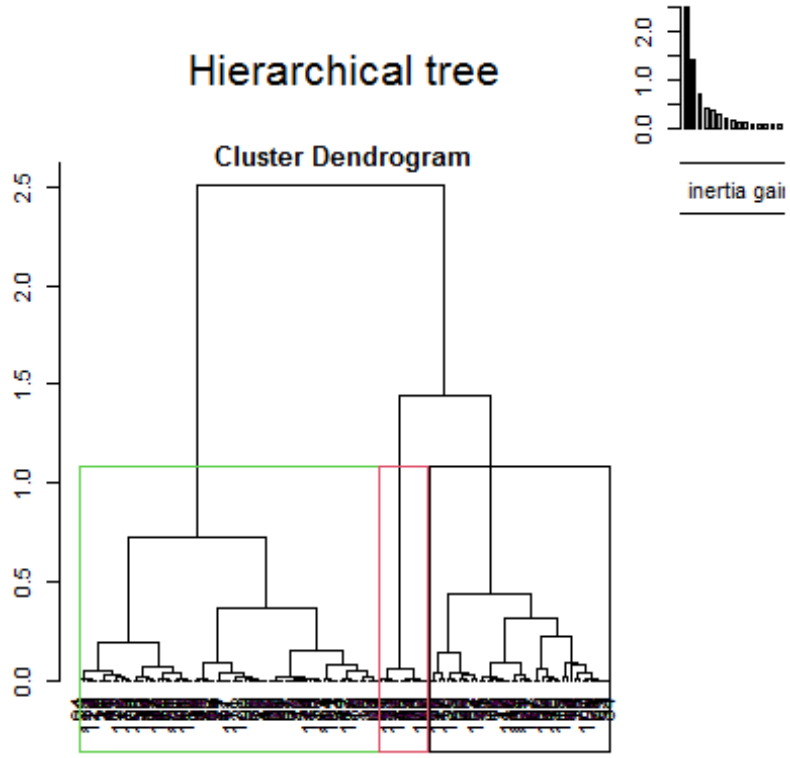
Les individus positionnés dans le quadrant supérieur droit sont ceus qui sont intéressés par "Plaisir de film" et "Plaisir Social" et les individus qui occupent le quadrant inférieur droit sont intéressés par "Eléments de production cinématographique".

3. Classification:

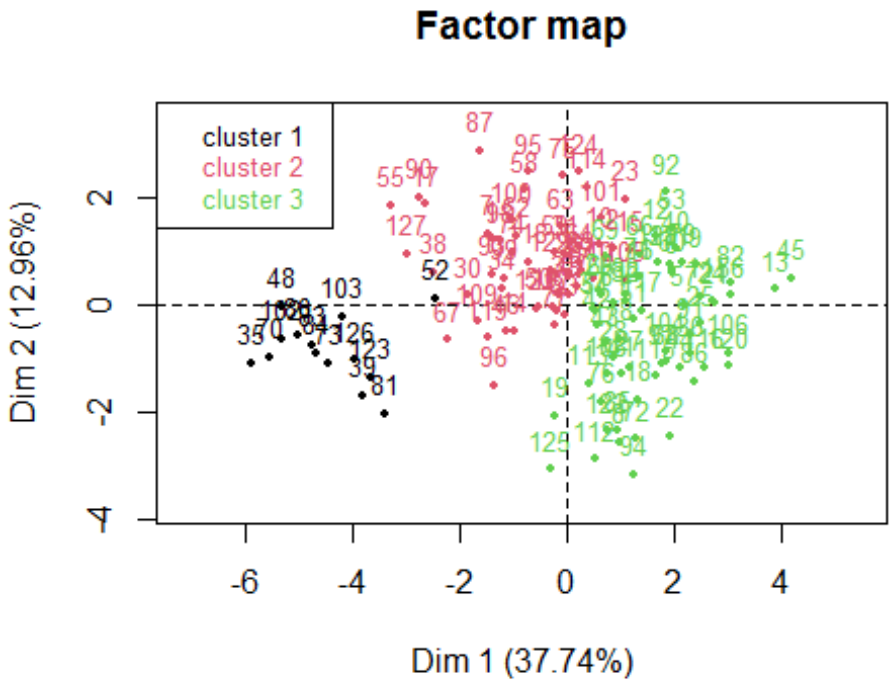
```
c<-dist(scale(reason),method="euclidean")
h<- hclust(c, method="ward.D2")
plot(h, hang = -1, cex =0.6)
```




```
res.HCPC<-HCPC(res.pca_R, consol=TRUE, graph=F)
plot.HCPC(res.HCPC,choice='tree', title ='Hierarchical tree')
```

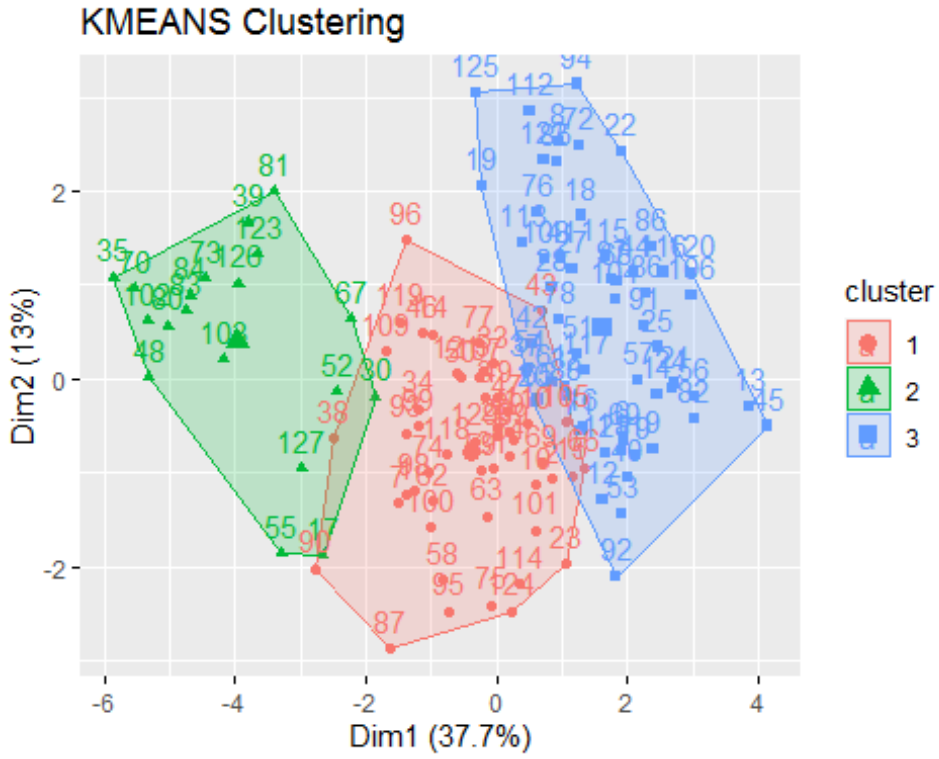



```
plot.HCPC(res.HCPC, choice = 'map', draw.tree=FALSE, title ='Factor map')
```



Methode de K_means:

```
res.km <- eclust(reason, "kmeans",k=3, nstart = 25)
```



Vérification de la validation de notre classification:

```
fviz_silhouette(res.km)
```

```
##   cluster size ave.sil.width
## 1       1   51          0.07
## 2       2   19          0.37
## 3       3   59          0.19
```

4. ACM

L'objectif de ce ACM est d'identifier : -Un groupe de personnes qui ont des réponses similaires aux questions posées. -Les associations entre les catégories des différentes variables.

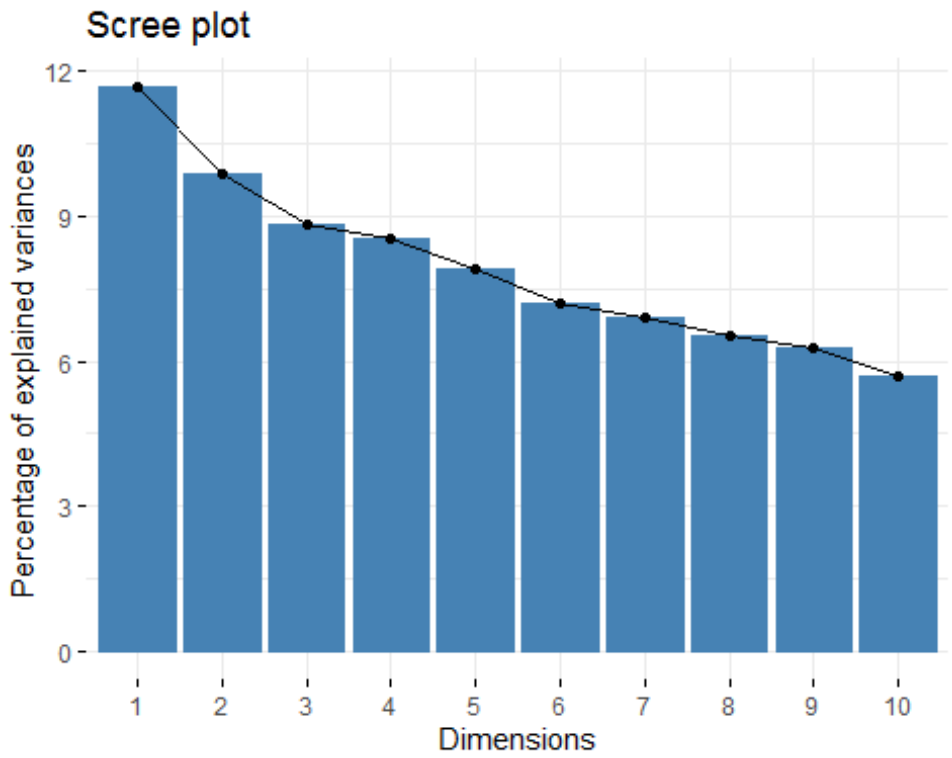
```
cinema$Catégorie_socioprofessionnelle<-
factor(cinema$Catégorie_socioprofessionnelle,c("Etudiant","Elève","Profession
libérale","Cadre et profession intellectuelle
supérieure","Retraité","Employé"))

tafc<-table(cinema$Catégorie_socioprofessionnelle,cinema$"Pour vous, quels
sont les critères d'une bonne salle de cinéma ?" )

M<-cinema[, c(2,34, 56,69)]
ncol(M)
```

```
## [1] 4
```

```
nrow(M)

## [1] 129
```

Choix de Nombre d'Axe:

```
res.mca <- MCA (M, graph = FALSE)
res.mca$eig

##         eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.43755857              11.668229                           11.66823
## dim 2  0.36992641               9.864704                           21.53293
## dim 3  0.33125876               8.833567                           30.36650
## dim 4  0.32021198               8.538986                           38.90549
## dim 5  0.29712233               7.923262                           46.82875
## dim 6  0.26909377               7.175834                           54.00458
## dim 7  0.25941401               6.917707                           60.92229
## dim 8  0.24523480               6.539595                           67.46188
## dim 9  0.23563997               6.283733                           73.74562
## dim 10 0.21296685               5.679116                           79.42473
## dim 11 0.19761133               5.269635                           84.69437
## dim 12 0.18033656               4.808975                           89.50334
## dim 13 0.16791620               4.477765                           93.98111
## dim 14 0.12791921               3.411179                           97.39229
## dim 15 0.09778925               2.607713                          100.00000

fviz_screeplot(res.mca)
```



Interpretation :

Critère de moyenne : puisque nbre de variable egala 4 , alors la moyenne sera 1/4=0.25 , de coup on Retenir les axes dont les valeurs propres sont supérieures à 0.25 donc les 7 premiers axes.

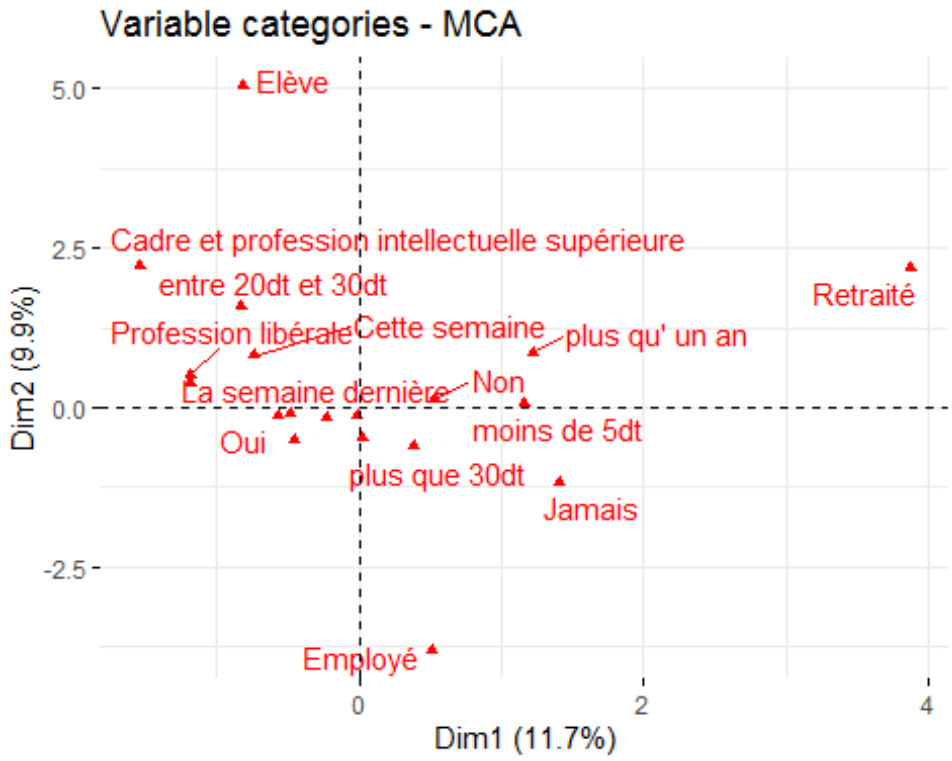
Critère de coude : le coude se trouve au niveau du 3éme axe.

Critère de taux d'inertie cumulé : on retient les 3 premiers axes .

=>En conclusion, le choix d'axe est les 3 premiers axes.

La première carte des modalités:

```
fviz_mca_var (res.mca,
              repel = TRUE,
              ggtheme = theme_minimal ())
```



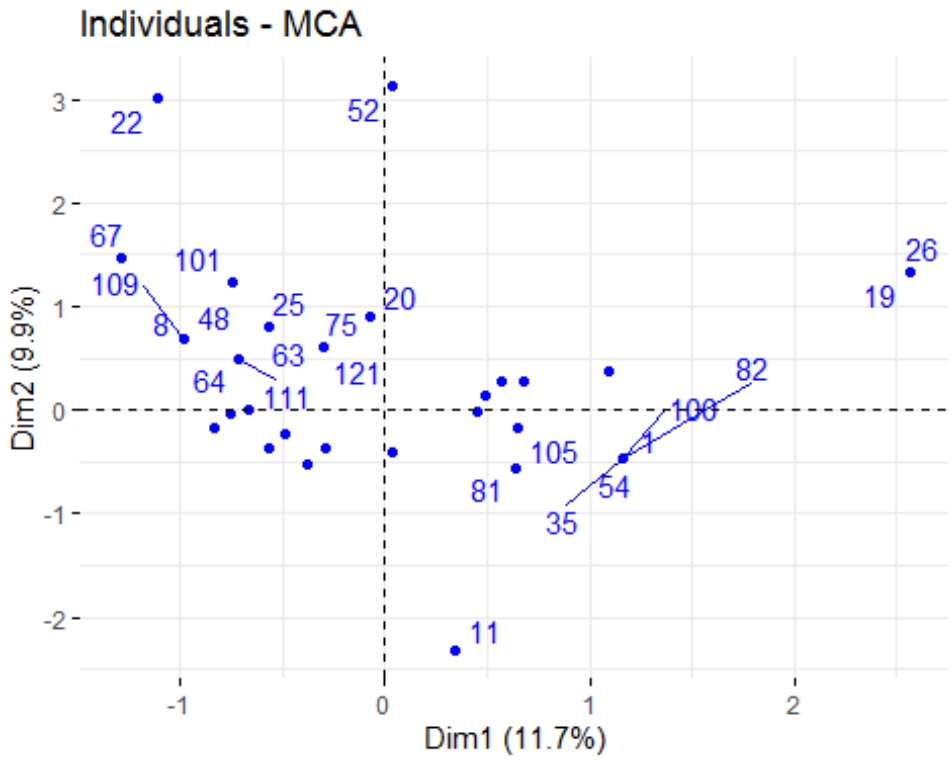
Interpretations:

-Nous remarquons que ceux qui habitent près d'un cinéma sont allés au cinéma au cours des 2 dernières semaines, tandis que ceux qui habitent loin des cinémas y sont allés beaucoup moins que les autres, en fait la majorité n'y sont jamais allés ou y sont allés une fois au cours de l'année précédente.

-On peut également constater que ceux qui travaillent en général dépensent en moyenne plus que les étudiants qui n'ont pas de revenu.

=>En général, on remarque que plus le cinéma est proche de chez soi, plus on a tendance à y aller. De plus, le revenu peut être corrélé à la fréquence à laquelle quelqu'un va au cinéma, ce qui est légitime car de nos jours, aller au cinéma peut être coûteux, surtout pour les étudiants. En effet, nous pouvons constater que ceux qui travaillent en général dépensent en moyenne plus que les étudiants qui n'ont pas de revenu.

```
fviz_mca_ind (res.mca,select.ind  = list(cos2 = 0.1),
              repel = TRUE,
              ggtheme = theme_minimal ())
```



Interpretation:

-Les individus positionnés dans le quadrant gauche sont ceus qui qui occupent des posts de travail et aller souvent au cinéma et les individus qui occupent le quadrant droit sont des etudiants qui allers pas fréquemment au cinéma et qui n'ont pas un revenu.