**RAPPORT DE**

**PROJET**

**D’ANALYSE DE DONNEES  :**

**3ème Année ESISA 2022-2023**

**SUJET :**

**Analyse de deux jeux de données.**

**ELABORE PAR :**

**MOHAMED-REDA BOUAMOUD**

**&**

**HOUDA BOUZOUBAA**

**ENCADRE PAR :**

**MR. ALAOUI AMINE**

1. Jeu de données 1: Augmentation de l’épaisseur de l’intima-media
   1. Statistiques descriptives :

Avant de commencer notre traitement, on trouve qu’ il est préférable de supprimer la variable ‘paqan’ vu qu’elle contient plusieurs données manquantes, et de se contenter de la variable ‘tabac’.

X<-read\_excel("Intima\_Media.xlsx")

View(X)

X<-X[,-6]

View(X)

str(X) #toutes les var sont quantitatives

summary(X)

75% des individus ont un âge <48 ans, une taille < 176 cm, un poids <75.75 kg , et la mesure de l’intima-media <0,57mm.

Nuages de point de la mesure d’entima-media en fonction de l’âge :

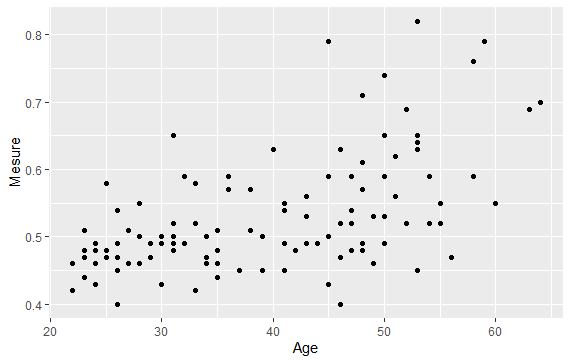
library(ggplot2)

ggplot()+

geom\_point(data = X, aes(x=AGE, y=mesure))+

xlab("Age")+

ylab("Mesure")



Selon ce nuage de point on note que l’épaisseur de l’intima-media dépend linéairement à l’âge de l’individu.

* 1. Nettoyage de données :

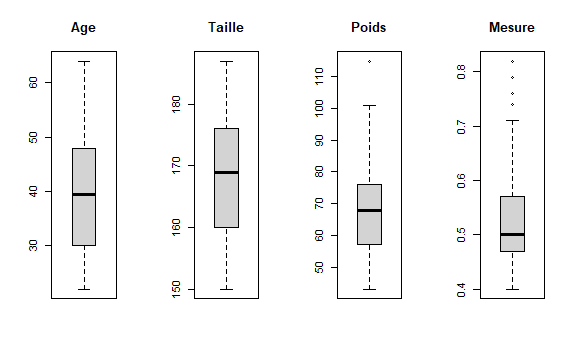
par(mfrow=c(1,4))

boxplot(X$AGE,main='Boite a moustache d age')

boxplot(X$taille,main='Boite a moustache de la taille')

boxplot(X$poids,main='Boite a moustache de poids')

boxplot(X$mesure,main='Boite a moustache de mesure')



Ces boites à moustache nous montre qu’il y a une seule donnée aberrante au niveau de la variable ‘poids’ qui dépasse 110 kg, et 4 données aberrantes au niveau de la variable ‘mesure’ dépassant 0.7 mm .

* 1. ACP :

Avant d’appliquer l’ACP on teste si les variables sont factorisables

On applique le test de KMO et le test de Bartlett :

C=cor(X)

C

KMO(C)

cortest.bartlett(C,n=110)

Msa = 0.68> 0.6 p-value=2.8\*10^(-30) <<0.05

Donc on peut appliquer l’ACP afin de réduire la dimension de notre modèle.

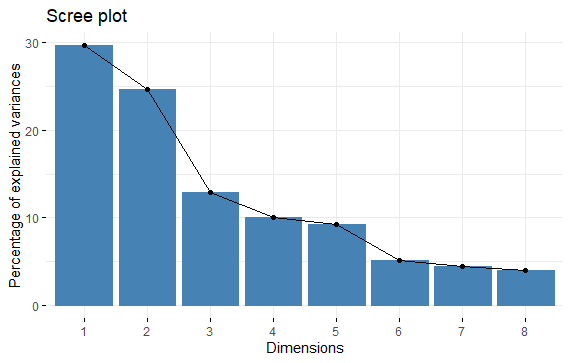
Choix du nombre d’axes :

pca = PCA(X, scale.unit = TRUE, graph = TRUE)

pca$eig

fviz\_eig(pca)

Le critère de Kaiser nous conduit à retenir 3 axes(1 ,2,3), cependant le critère du coude nous montre une chute importante à partir du 3 -ème axe (du 25% a 13%) donc on garde les 3 axes (1,2,3).



On relance l’ACP avec 3 axes seulement

pca = PCA(X, scale.unit = TRUE,ncp = 3, graph = TRUE)

summary(pca)

Variables

Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2

SEXE | -0.848 30.321 0.719 | 0.065 0.217 0.004 | 0.055 0.296 0.003 |

AGE | -0.289 3.522 0.084 | 0.836 35.510 0.699 | 0.012 0.015 0.000 |

taille | 0.863 31.429 0.746 | -0.185 1.733 0.034 | -0.067 0.433 0.004 |

poids | 0.776 25.403 0.603 | 0.212 2.273 0.045 | -0.249 6.013 0.062 |

tabac | 0.212 1.891 0.045 | 0.509 13.170 0.259 | 0.400 15.543 0.160 |

SPORT | 0.173 1.258 0.030 | -0.223 2.529 0.050 | 0.871 73.524 0.758 |

mesure | 0.043 0.078 0.002 | 0.794 32.056 0.631 | -0.094 0.863 0.009 |

alcool | 0.380 6.098 0.145 | 0.496 12.512 0.246 | 0.185 3.314 0.034 |

Individuals (the 10 first)

Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2

1 | 1.629 | 0.847 0.275 0.270 | -0.056 0.001 0.001 | -0.613 0.331 0.142 |

2 | 2.836 | 0.193 0.014 0.005 | -0.287 0.038 0.010 | 0.113 0.011 0.002 |

3 | 2.897 | -1.590 0.969 0.301 | 1.371 0.868 0.224 | -0.759 0.508 0.069 |

4 | 1.822 | -0.028 0.000 0.000 | 0.019 0.000 0.000 | 1.171 1.209 0.413 |

5 | 2.962 | -2.469 2.336 0.695 | 0.259 0.031 0.008 | -0.491 0.212 0.027 |

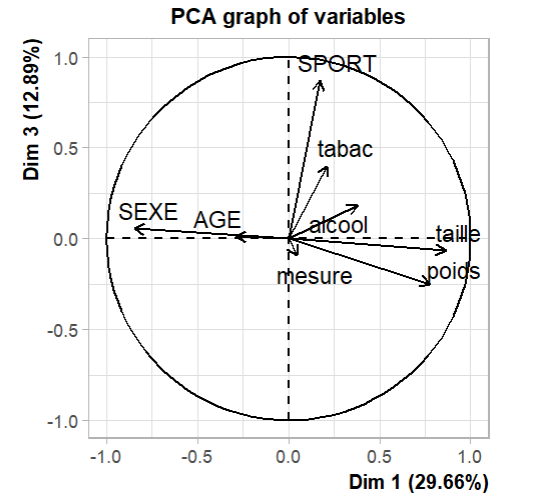
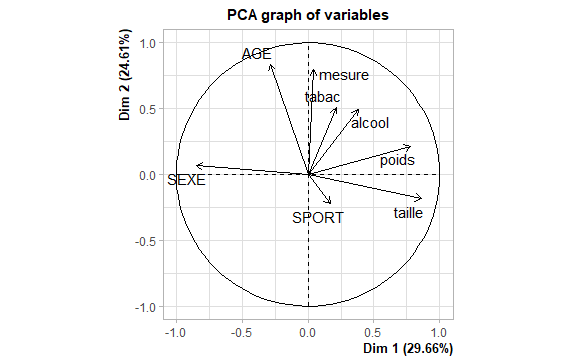
6 | 2.852 | -1.495 0.856 0.275 | 1.136 0.596 0.159 | 0.419 0.154 0.022 |

7 | 2.950 | 1.782 1.217 0.365 | -2.217 2.270 0.565 | 0.591 0.308 0.040 |

8 | 2.377 | -0.751 0.216 0.100 | -1.107 0.566 0.217 | 1.510 2.011 0.404 |

9 | 3.040 | 1.220 0.571 0.161 | 1.956 1.767 0.414 | -1.114 1.094 0.134 |

10 | 1.951 | -0.077 0.002 0.002 | 0.247 0.028 0.016 | -1.103 1.073 0.320 |

Les graphes obtenus:

Dans le plan(1,2) l’âge et le sexe sont les variables qui participent le plus à la formation de l’axe par contre on peut voir que la variable sport est bien projetée sur le plan(1,3).

On voit aussi qu’ il y a 2 groupes de variables opposées : celles qui contribuent positivement( taille, poids …)à l’axe et celles qui contribuent négativement(sexe, âge).

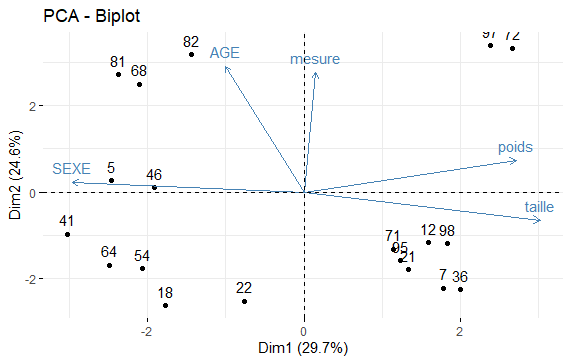
Choix des individus :

Vu que la plupart des individus ont une contribution > 1/110 , alors on met le seuil de qualité = 0.7

fviz\_pca\_ind(pca,select.ind=list(indiv$contrib>1/110))

fviz\_pca\_ind(pca,select.ind=list(cos2=0.7)) #qualite >0.7

fviz\_pca\_biplot(pca,select.ind=list(cos2=0.7),select.var=list(cos2=0.5))



Les individus (81, 82…) se caractérisent par un grand âge contrairement au individus (7, 36… ) qui se caractérisent par leur grand poids, et on note aussi que le sexe masculin a une mesure d’intima media marquée que le sexe féminin.

* 1. Régression :

Vu que les variables sont toutes quantitatives, la régression adéquates est la régression linéaire.

resreg=lm(mesure~ SEXE + AGE + taille + poids + tabac +SPORT + alcool,data=X)

resreg

summary(resreg)

p-value: 2.553e-08< 0.05 Adjusted R-squared: 0.33

Ce modèle est non constant, n’explique pas bien les données c-v-dire les données sont pas bien ajustées sur la droite de regression.

m0=lm(mesure~1,data=X)#model constant

AIC(m0)

BIC(m0)

summary(m0)

#demarches but; les var les plus pertinants

#demarche descendante

B=step(resreg,direction="backward")#complet

#AGE poids

summary(B)

AIC(B)

BIC(B)

#demarche ascendante

F=step(m0,scope=list(lower=m0, upper=resreg),direction="forward")

summary(F)

AIC(F)

BIC(F)

#demarch mixte

M=step(m0,scope=list(upper=resreg),direction="both")

summary(M)

AIC(M)

BIC(M)

En appliquant les 3 démarches ascendantes, descendantes et mixtes on obtient le même model avec les variables **AGE** et **poids.**

Conclusion :

L’âge élevé, le sexe masculin et le surpoids sont les facteurs déterminants de la majoration de l’épaisseur intima-média en revanche le tabagisme ne l’affecte pas significativement.

1. Jeu de données 2 : (l’évolution du chiffre d’affaires)
   1. Statistiques descriptives :

X<-read\_excel("Classeur2.xlsx")

View(X)

str(X)

X=X[-length(X$Y),-1]

Y=as.integer(X$Y)

Y

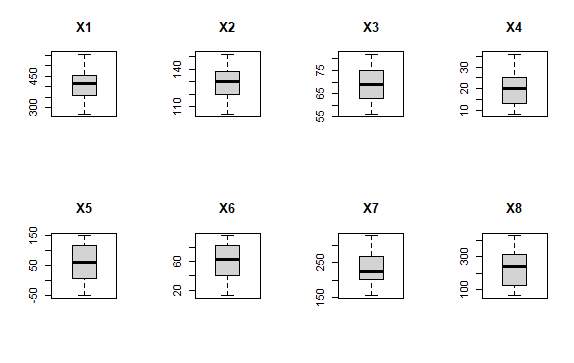
X=X[,-9]

View(X)

summary(X)

75% des individus ont X1<451, X2<138, X3<74.75, X4<24.75, X5<116, X6<81.5, X7<263.8, X8<313 .

* 1. Nettoyage de données :



Les boites à moustaches nous montres qu’il n y a pas de données aberrantes.

* 1. ACP :

Avant d’appliquer l’ACP on test si les données sont factorisables

C=cor(X)

dim(X)

C

KMO(C)

cortest.bartlett(C,n=38)

MSA = 0.55 <0.6

MSA for each item =

X1 X2 X3 X4 X5 X6 X7 X8

0.54 0.51 0.70 0.50 0.33 0.44 0.55 0.53

X5, X6, X2 ont un MSA < 0.6

P-value= 6.067444e-11<<0.05

Donc on peut appliquer l’ACP

Choix du nombre d’axes :

pca$eig

eigenvalue percentage of variance cumulative percentage of variance

comp 1 | 2.74097011 34.2621263 34.26213

comp 2 | 1.26095621 15.7619526 50.02408

comp 3 | 1.20130084 15.0162605 65.04034

comp 4 | 0.88024257 11.0030321 76.04337

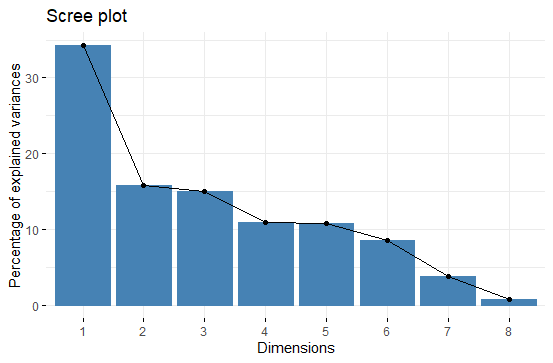
comp 5 | 0.86815886 10.8519857 86.89536

comp 6 | 0.68061102 8.5076378 95.40300

comp 7 | 0.30234134 3.7792668 99.18226

comp 8 | 0.06541906 0.8177382 100.00000

fviz\_eig(pca)



Le critère de Kaiser nous conduit à retenir 3 axes(1 ,2,3), cependant le critère du coude nous montre une chute importante à partir du 2 -ème axe (du 35% a 16%) donc on garde les 3 axes (1,2,3).

On relance l’ACP avec 3 axes seulement

pca = PCA(X, scale.unit = TRUE,ncp = 3, graph = TRUE)

summary(pca)

Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2

X1 | 0.846 26.106 0.716 | 0.288 6.560 0.083 | 0.138 1.586 0.019 |

X2 | -0.071 0.182 0.005 | -0.021 0.036 0.000 | -0.773 49.770 0.598 |

X3 | 0.801 23.417 0.642 | 0.114 1.033 0.013 | -0.274 6.264 0.075 |

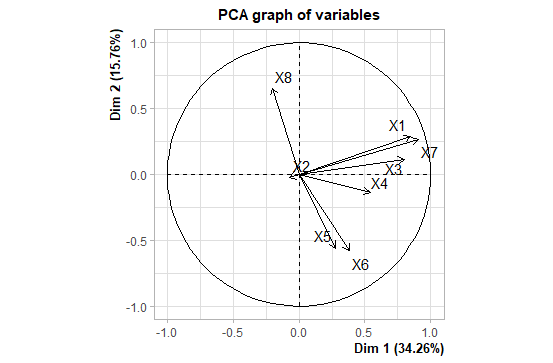
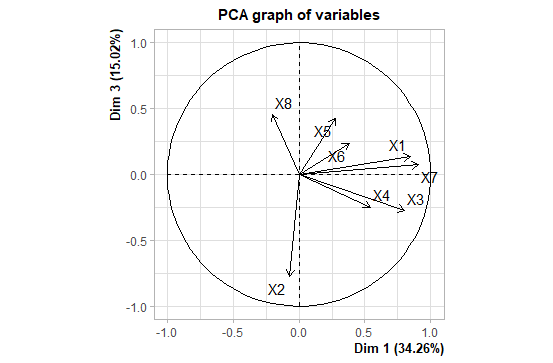
X4 | 0.541 10.678 0.293 | -0.136 1.462 0.018 | -0.248 5.119 0.061 |

X5 | 0.276 2.775 0.076 | -0.560 24.837 0.313 | 0.425 15.039 0.181 |

X6 | 0.382 5.331 0.146 | -0.579 26.616 0.336 | 0.236 4.617 0.055 |

X7 | 0.907 30.042 0.823 | 0.269 5.729 0.072 | 0.073 0.446 0.005 |

X8 | -0.201 1.469 0.040 | 0.652 33.728 0.425 | 0.454 17.158 0.206 |



Dans le plan(1,2) X7, X1 et X3 sont les variables qui participent le plus à la formation de l’axe par contre on peut voir que la variable X2 est bien projetée sur le plan(1,3).

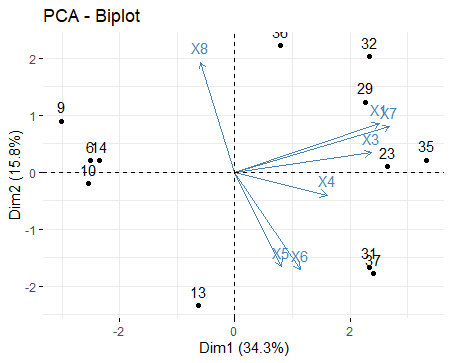
On voit aussi qu’ il y a 2 groupes de variables opposées : celles qui contribuent positivement( X6,X5 …)à l’axe et celles qui contribuent négativement(X8, X2).

Choix des individus :

Vu que la plupart des individus ont une contribution > 1/38 , alors on met le seuil de qualité = 0.6

fviz\_pca\_ind(pca,select.ind=list(cos2=0.6))

fviz\_pca\_biplot(pca,select.ind=list(cos2=0.6),select.var=list(cos2=0.5))



Les individus 29, 23… se caractérisent par X1, X7, X3

* 1. Régression :

Vu que les variables sont toutes quantitatives, la régression adéquates est la régression linéaire.

resreg=lm(Y~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,data=X)

resreg

summary(resreg)

Adjusted R-squared: 0.7522

Proche de 1 le model représente bien les données.

p-value: 1.957e-08 << 0.05

Le model est non constant.

m0=lm(y~1,data=X)#model constant

AIC(m0)

BIC(m0)

summary(m0)

#demarche descendante

B=step(resreg,direction="backward")#complet

#X1 X3 X5 X6

summary(B)

AIC(B)

BIC(B)

#demarche ascendante

F=step(m0,scope=list(lower=m0, upper=resreg),direction="forward")

summary(F)

AIC(F)

BIC(F)

#demarch mixte

M=step(m0,scope=list(upper=resreg),direction="both")

summary(M)

AIC(M)

BIC(M)

En appliquant les 3 démarches ascendantes, descendantes et mixtes on obtient le même model avec les variables X1, X3, X5 et X6.

En comparant les modèles obtenus avec les critères AIC et BIC on trouve que le meilleur parmi eux est le modèle obtenu par les 3 démarches.

La prédiction du chiffre d’affaires Y dans le semestre 39 :

X$Y[39] = 0

X$Y = as.integer(X$Y)

View(X)

resreg=lm(Y~ X1 + X3 + X5 + X6 ,data=X)

Summary(resreg)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5354.873 1379.965 3.880 0.000456 \*\*\*

X1 4.252 2.680 1.587 0.121853

X3 -36.132 24.781 -1.458 0.154006

X5 2.605 2.618 0.995 0.326642

X6 4.009 6.291 0.637 0.528212

pre= predict(resreg)#y^

pre

y= 4972.780.

Conclusion :

Les facteurs augmentant le chiffre d’affaires sont X1, X5, X6, en revanche X3 est un facteur qui diminue le chiffre d’affaires.