Rapport de Projet

Analyse de données

Réalisé par : *Bouibauan mohamed*

Maataoui Mohamed

2021-2022

Table des matières

1 Chapitre 1 : Régression linéaire multiple

1.1 Le jeu de données

X	turda	X2.house.age	X3.distance.to.the.nearest.MRT.station	X4.number.of.convenience.stores	X5.latitude	X6.longitude	dephm	SOLID	Condi	Y.house.price.of.unit.area
0	1604.093	32.0	84.87882	10	24.98298	121.5402	0	2912.211	532.8852	37.9
1	1603.160	19.5	306.59470	9	24.98034	121.5395	8	8896.195	430.8368	42.2
2	1603.993	13.3	561.98450	5	24.98746	121.5439	10	20988.429	381.3573	47.3
3	1604.064	13.3	561.98450	5	24.98746	121.5439	19	21144.975	306.3559	54.8
4	1602.545	5.0	390.56840	5	24.97937	121.5425	20	22372.303	341.7602	43.1
5	1603.357	7.1	2175.03000	3	24.96305	121.5125	30	14859.060	445.0375	32.1
6	1604.181	34.5	623.47310	7	24.97933	121.5364	39	16905.802	481.3071	40.3
7	1603.783	20.3	287.60250	6	24.98042	121.5423	50	14462.674	534.8010	46.7
8	1602.283	31.7	5512.03800	1	24.95095	121.4846	58	2552.963	517.4275	18.8
9	1604.193	17.9	1783.18000	3	24.96731	121.5149	75	7013.212	419.7889	22.1
10	1603.878	34.8	405.21340	1	24.97349	121.5337	78	10859.554	358.0563	41.4
11	1603.354	6.3	90.45606	9	24.97433	121.5431	100	34226.072	415.5775	58.1
12	1602.752	13.0	492.23130	5	24.96515	121.5374	117	15249.620	361.9016	39.3
13	1602.856	20.4	2469.64500	4	24.96108	121.5105	125	18409.037	389.6832	23.8
14	1602.957	13.2	1164.83800	4	24.99156	121.5341	150	26132.212	326.0262	34.3
15	1603.947	35.7	579.20830	2	24.98240	121.5462	152	28001.118	365.0916	50.5
16	1604.532	0.0	292.99780	6	24.97744	121.5446	200	21575.245	444.1166	70.1
17	1604.482	17.7	350.85150	1	24.97544	121.5312	221	15335.511	399.6054	37.4
18	1604.649	16.9	368.13630	8	24.96750	121.5445	250	29542.338	456.3085	42.3

FIGURE 2

——>le but de ces données est la prédiction du prix des logements dans la banlieue de Boston..

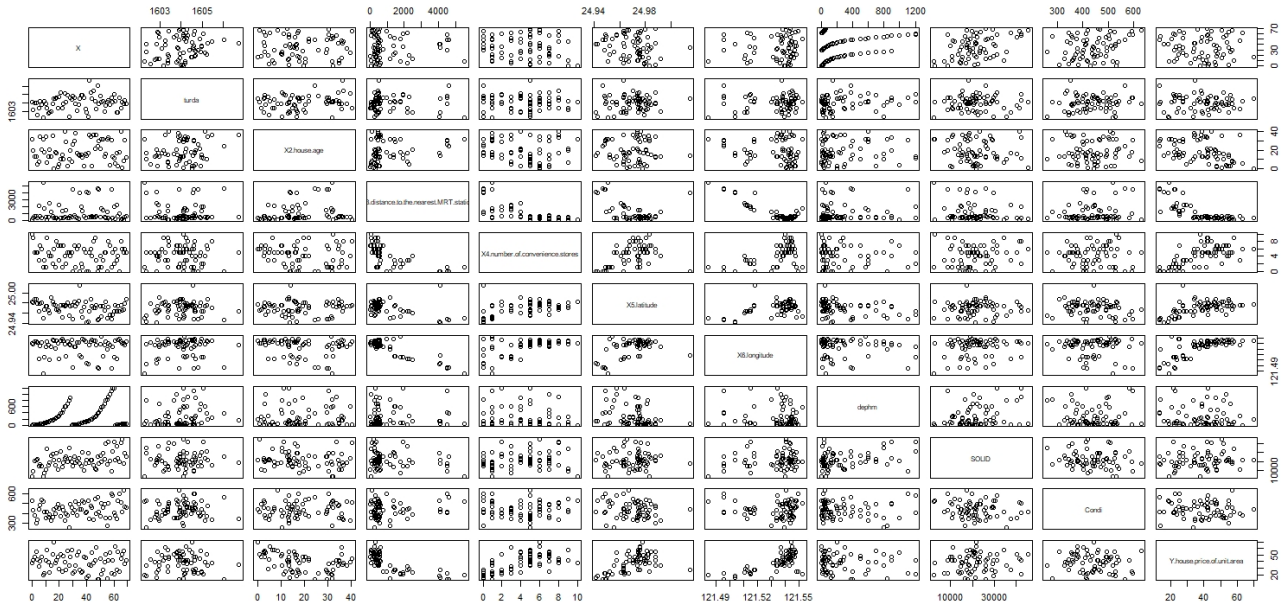


FIGURE 3

1.2 Modèle de régression linéaire multiple incluant toute les variables

```
#model de regression lineaire
modele <- lm(Y.house.price.of.unit.area~.,data=data)
```

1.2.1 les variables explicatives non significatives

```
> summary(modele)$coefficients
                                         Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)                          -1.780052e+04 1.752935e+04 -1.0154691 0.3139597143
turda                                -4.567688e-01 1.212037e+00 -0.3768604 0.7076069028
X2.house.age                         -3.263998e-01 9.210312e-02 -3.5438520 0.0007708619
X3.distance.to.the.nearest.MRT.station -3.448119e-03 1.988284e-03 -1.7342183 0.0880148172
X4.number.of.convenience.stores       9.729484e-01 5.315990e-01  1.8302300 0.0721837845
X5.latitude                           2.310774e+02 9.656375e+01  2.3930038 0.0198574041
X6.longitude                          1.053800e+02 1.459596e+02  0.7219804 0.4731109996
dephm                                 4.567810e-04 3.460764e-03  0.1319885 0.8954350286
SOLID                                -6.085670e-05 1.166753e-04 -0.5215901 0.6038759014
Condi                                 3.739665e-03 1.346177e-02  0.2777988 0.7821217913
```

La constant et **les variables** Turda,X3.distance.to.the.nearest.MRT.station, X6.longitude,dephm,SOLID,Condi sont significativement nul car ils ont une Valeur de **P-Value >5 %**

1.2.2 la valeur de R^2 et R^2_{ajus}

-Multiple R-squared: 0.6783,

-Adjusted R-squared: 0.6301

1.2.3 Le test de Fisher et sa signification

F-statistic : 14.06 >0 on 9 and 60 DF
P-value : 7.39e-12 <5% donc est Valide
==> l'hypothèse H0 qui dit que tous les coefficients sont nuls est rejetée car p-value de Fisher <5

1.3 Amélioration du modèle initiale par la procédure de step

-la fonction step élimine a chaque fois une variable on commence par l'introduction de tout les variable c'est a dire tout "V"
-SI une variable est éliminer on met " F" dans la case sélectionnées

Step1 :Start : AIC=305.07

Variables	Variables sélectionnées
turda	V
X2.house.age	V
X3.distance.to.the.nearest.MRT.station	V
X4.number.of.convenience.stores	V
X5.latitude	V
X6.longitude	V
dephm	V
SOliD	V
Condi	V

Step2 : AIC =303.09

Variables	Variables sélectionnées
turda	V
X2.house.age	V
X3.distance.to.the.nearest.MRT.station	V
X4.number.of.convenience.stores	V
X5.latitude	V
X6.longitude	V
dephm	F
SOLID	V
Condi	V

Step 3 : AIC=301.18

Variables	Variables sélectionnées
turda	V
X2.house.age	V
X3.distance.to.the.nearest.MRT.station	V
X4.number.of.convenience.stores	V
X5.latitude	V
X6.longitude	V
dephm	F
SOUD	V
Condi	F

Step 4 : AIC=299.31

Variables	Variables sélectionnées
turda	F
X2.house.age	V
X3.distance.to.the.nearest.MRT.station	V
X4.number.of.convenience.stores	V
X5.latitude	V
X6.longitude	V
dephm	F
SOUD	V
Condi	F

Step 5 : AIC=297.67

Variables	Variables sélectionnées
turda	F
X2.house.age	F
X3.distance.to.the.nearest.MRT.station	V
X4.number.of.convenience.stores	V
X5.latitude	V
X6.longitude	V
dephm	F
SOUD	F
Condi	F

Step 6 : AIC=296.38 **Fin**

Variables	Variables sélectionnées
turda	F
X2.house.age	F
X3.distance.to.the.nearest.MRT.station	V
X4.number.of.convenience.stores	V
X5.latitude	V
X6.longitude	F
dephm	F
SOUD	F
Condi	F

1.3.1 Remarques :

-la fonction step élimine a chaque fois une variable
-la diminuation du critière AIC à chaque step

-Le Modèle choisi par step :

```
Call:
lm(formula = Y.house.price.of.unit.area ~ X2.house.age + X3.distance.to.the.nearest.MRT.station +
    X4.number.of.convenience.stores + X5.latitude, data = data)

Coefficients:
                          (Intercept)                  X2.house.age  X3.distance.to.the.nearest.MRT.station
                            -5.810e+03                    -3.364e-01                            -4.459e-03
X4.number.of.convenience.stores                  X5.latitude
                            9.435e-01                     2.345e+02
```

Multiple R-squared : 0.6723 **Adjusted R-squared : 0.6521**

Conclusion :

-Grace au fonction **step** Adjusted R-squared agumenter de **0.6301** à **0.6521** donc ila ya une amélioration .

-Multiple R-squared diminué mais c'est pas un critère car les deux modèle n'a pas les même nombre de variables

1.3.2 les tests de validation :

-Est-ce que le modèle représente bien la réalité ?
==>Ona plusieur critère :

-Test d'himosadicite
-Test de Normalité
-Les valeurs aberrantes

1.3.3 Test d'himosadicite

-On remarque que la distribution des résidus en fonction des predictions est n'est pas aléatoire alors on peut dire l'homoscédasticité n'est pas vérifiée.

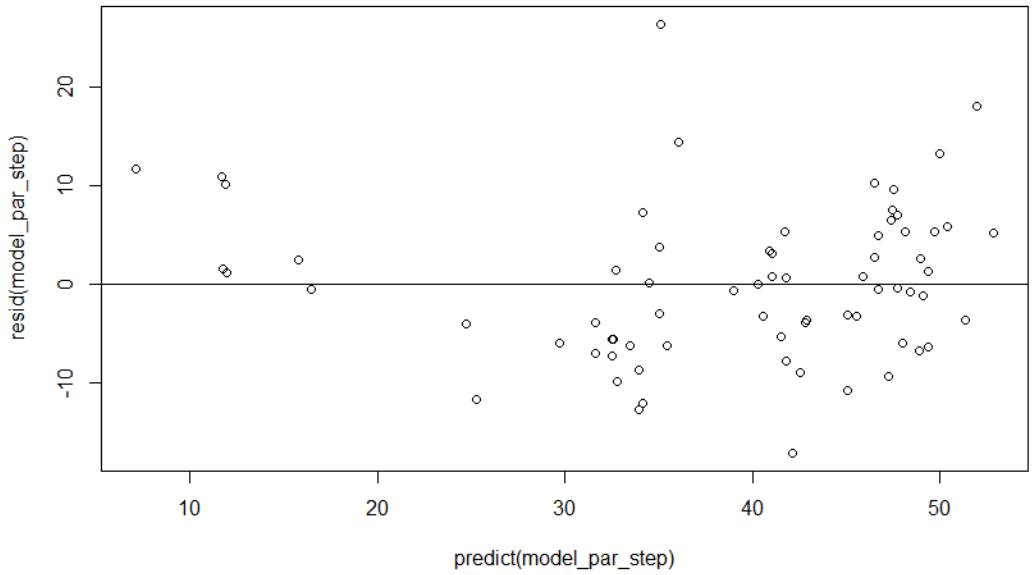


FIGURE 4

1.3.4 Test de Normalité

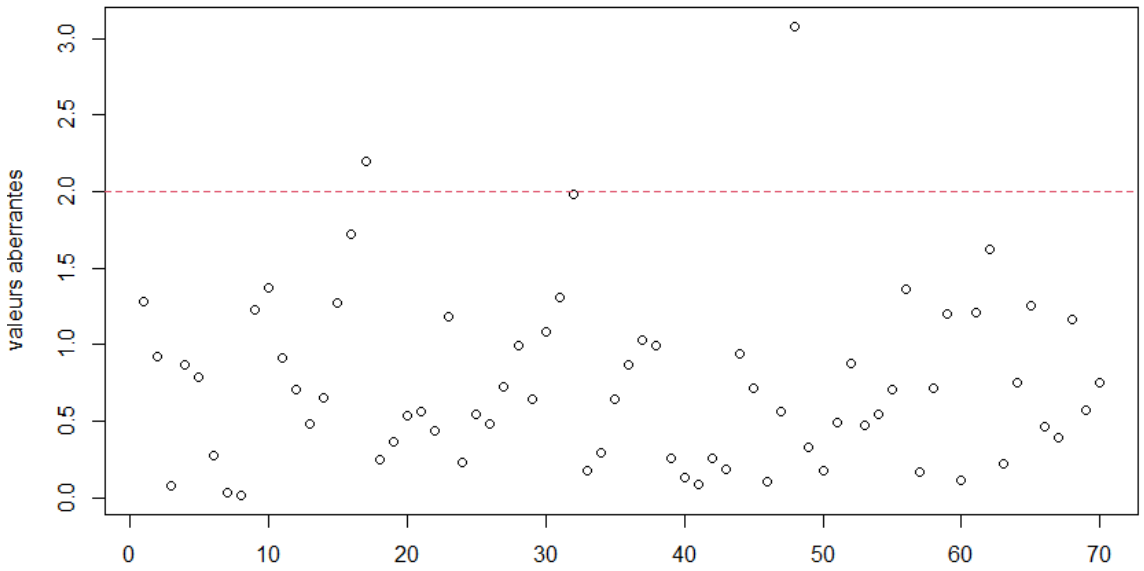
-Test de Kolmogorov-Smirnov:

p-value = 4.386e-12 ===> la normalité des résidus est rejeté par KS

-Test de Shapiro:

p-value = 0.2273 ===> la normalité des résidus est acceptée par Shapiro

1.3.5 Les valeurs aberrantes



1.4 la méthode pas à pas de sélection des variables

les critières de selection :

A chaque étape :
La var entrante est celle qui présente le plus grand F avec pvalue<10%
La var sortante est celle qui présente le plus petit F avec pvalue>10%

Arrêt :
Si les var entrantes ont p-values>10% et les var sortantes ont des pvalues<10%

$$pvalue = P(F(1, n-k-2) > F)$$

Les etapes qui on a suivre :

-1) On commence la méthode pas à pas par l'intégration de la variable la plus significative (F le plus grand)
c'est la variable**X3.distance.to.the.nearest.MRT.station** dans notre cas F=77.88 et P-value=6.981082e-13<10% :

X	turda	X2.house.age	X3.distance.to.the.nearest.MRT.station	X4.number.of.convenience.stores	X5.latitude	X6.longitude	dephm	SOLID	Condi

```
Fish
[1]  0.0000000  0.1245232 11.8975393 77.8864991 44.0265183 28.5210545 62.7156250  0.9442940  0.7587439  0.5627993
```

-2) introduction de variable x2.house.age
F=12.65 et P-value=0.0006949426<10% :

X	turda	X2.house.age	X3.distance.to.the.nearest.MRT.station	X4.number.of.convenience.stores	X5.latitude	X6.longitude	dephm	SOLID	Condi

```
> Fish
 [1]  0.0000000  0.8152433 12.6500209  0.0000000  4.8052973  6.2335269  1.1346776  1.6375342  0.1622141  0.1805225
```

-3) Aucune variable n'est retirée, les F sont significatifs

-4) introduction de variable X5.latitude
F= 8.86 et P-value=0.0040<10% :

X	turda	X2.house.age	X3.distance.to.the.nearest.MRT.station	X4.number.of.convenience.stores	X5.latitude	X6.longitude	dephm	SOLID	Condi

```
Fish
1] 0.00000000 0.05439503 0.00000000 0.00000000 4.99954449 8.86128440 0.62335340 1.48047389 0.58170436 0.18675176
```

-5) introduction de variable X4.number.of.convenience.stores
F= 3.56 et P-value= 0.0634<10% :

X	turda	X2.house.age	X3.distance.to.the.nearest.MRT.station	X4.number.of.convenience.stores	X5.latitude	X6.longitude	dephm	SOLID	Condi

```
Fish
[1] 0.00000000 0.14311502 0.00000000 0.00000000 3.56529304 0.00000000 0.24393364 0.09967003 0.37992824 0.53867054
```

-6) Aucune variable n'est retirée, les F sont significatifs
-on a calculer le Fisher pour ajouter un variable signifactif mais ona a touver que P-value du variable qui a plus grand F est égal à **0.4244>10%** donc aucun variable à entrer ou à sortie d'après 6
==>le test d'arrêt est vérifier.

Tableau de resultats du Méthode pas à pas :

- F et P-value dans le tableau c'est du modèle avec les variables choisi à chaque etape
- fisher avec p-value du critère de selection du variable entrant et sortant et déja citer en haut "les etapes qui on a suivre"

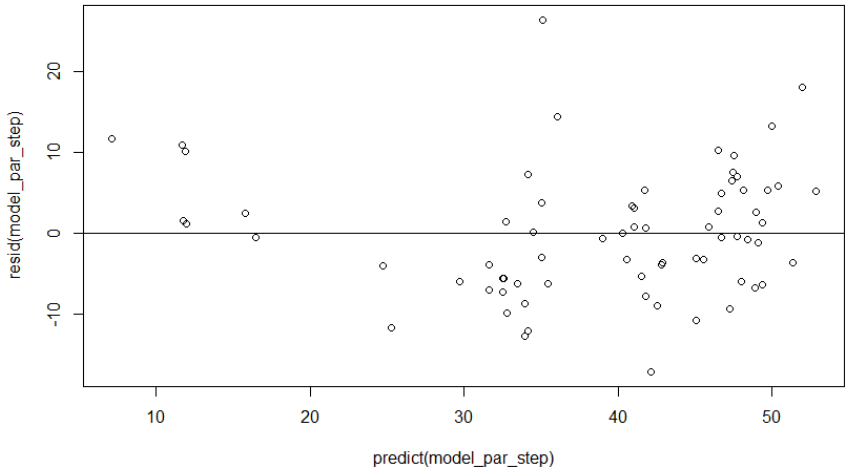
Etape	Var entrée	Var sortie	R^2_{ajus}	F	P-value
0	X3.distance.to.the.nearest.MRT.station	aucun	0.527	77.89	$6.981e-13$
1	X2.house.age	aucun	0.5962	51.94	$2.39e-14$
2	X5.latitude"	aucun	0.6386	41.64	$3.191e-15$
3	X4.number.of.convenience.stores	aucun	0.6521	33.34	$4.092e-15$

1.4.1 les tests de validation :

-Puisque nous avons trouvé le même modèle que celui sélectionné par la procédure step ,nous avons donc les mêmes tests de validation

1.4.2 Test d'himosadicite

-On remarque que la distribution des résidus en fonction des predictions est n'est pas aléatoire alors on peut dire l'homoscédasticité n'est pas vérifiée.



1.4.3 Test de Normalité

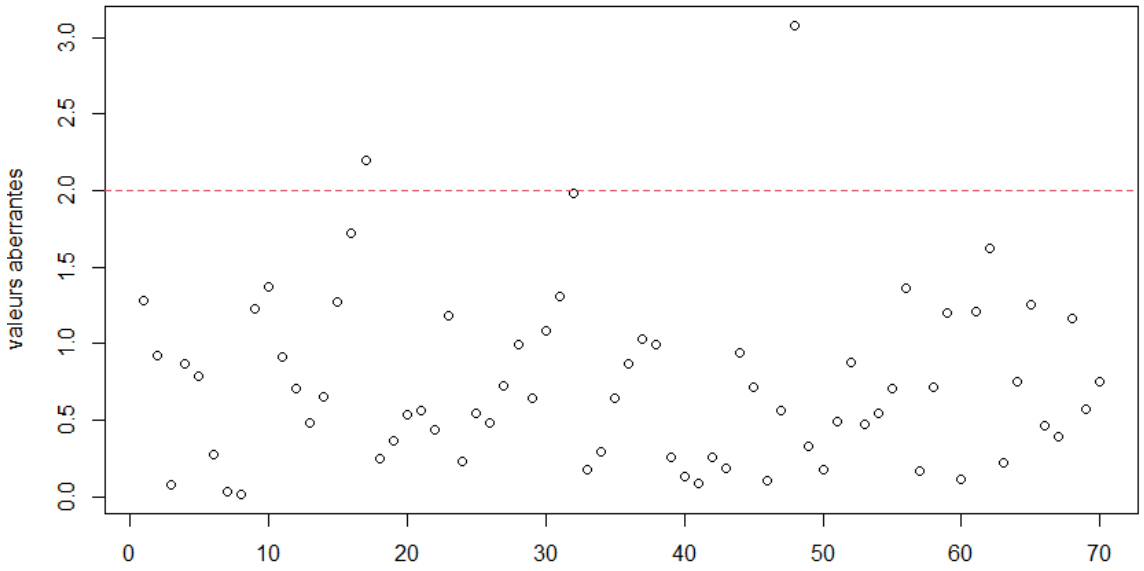
-Test de Kolmogorov-Smirnov:

p-value = 4.386e-12 ===> la normalité des résidus est rejeté par KS

-Test de Shapiro:

p-value = 0.2273 ===> la normalité des résidus est acceptée par Shapiro

1.4.4 Les valeurs aberrantes



1.4.5 le critère AIC du modèle obtenu

-le modèle sélectionner par les deux Méthodes :
model<-lm(Y.house.price.of.unit.area X2.house.age + X3.distance.to.+ X4.number.of.convenience.stores + X5.latitude, data = data)

```
AIC(model_par_step)
  497.0287
AIC(model_par_fischer)
  497.0287
```

2 Conclusion

-Dans notre cas les deuxméthodes de selection des variables donne les mèmes Resultats

3 Chapitre 2 :Les méthodes de classification

3.1 kmeans

K-means est un algorithme de clustering.,Il consiste à regrouper les éléments de notre jeu de donnée en groupes, appelés clusters. Le but est de faire ressortir les patterns cachés dans la donnée en regroupant les éléments qui se « ressemblent ».

L'algorithme des k-moyens regroupe les points en k clusters. Cela suppose qu'il faut avoir une idée du nombre de clusters pour appliquer cet algorithme.

3.1.1 Description de données :

```
> str(data)
'data.frame':   70 obs. of  11 variables:
 $ X                                     : int  0 1 2 3 4 5 6 7 8 9 ...
 $ turda                                 : num  1604 1603 1604 1604 1603 ...
 $ X2.house.age                          : num  32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
 $ X3.distance.to.the.nearest.MRT.station: num  84.9 306.6 562 562 390.6 ...
 $ X4.number.of.convenience.stores       : int  10 9 5 5 5 3 7 6 1 3 ...
 $ X5.latitude                           : num  25 25 25 25 25 ...
 $ X6.longitude                          : num  122 122 122 122 122 ...
 $ dephm                                 : num  0 8 10 19 20 30 39 50 58 75 ...
 $ SOLID                                 : num  2912 8896 20988 21145 22372 ...
 $ Condi                                 : num  533 431 381 306 342 ...
 $ Y.house.price.of.unit.area            : num  37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1 ...
```

3.1.2 Normalisation

-Cette operation permet de centré et de réduire les données en divisant sur l'ecart type non biaisé pour accélerer l'operation de processing des données

3.2 Application de kmeans

-on commence par nombre de groupes **center = 6** et **nstart = 5** nombre d'essais avec différents individus de départ, c'est-à-dire 5 exécution de kemeans avec différents choix de classe initiale .
groupes.kmeans <- kmeans (data.actifs,centers =6, nstart =5)
|

3.2.1 affichage des résultats :

```
> print(groupes.kmeans)
K-means clustering with 6 clusters of sizes 12, 17, 7, 13, 9, 12

Cluster means:
        turda X2.house.age X3.distance.to.the.nearest.MRT.station X4.number.of.convenience.stores X5.latitude X6.longitude
1  0.89001090    1.0318206                          -0.4873911                       0.6461264   0.4311625    0.4764639
2  0.07417499    0.2823709                           0.2967160                      -0.9589810  -0.1959471   -0.2928688
3 -0.23954990    0.5097803                           2.5145266                      -1.4364298  -1.7765055   -2.2614861
4 -0.10506315   -0.6693349                          -0.6039875                       0.8473991   0.3987236    0.6585649
5 -0.06296237   -0.4406623                          -0.4543341                       0.3138667  -0.1966947    0.2074337
6 -0.69431451   -0.6736084                          -0.4046934                       0.3969316   0.5982946    0.3886132
        dephm       SOLID       Condi
1 -0.20021207  0.4217135 -0.33346559
2 -0.07631418 -0.4218050  0.02877736
3  0.10220631 -0.4899475  0.66466056
4 -0.45399506 -0.6589658  0.75814040
5  1.91519310  0.9880881 -0.17540007
6 -0.69586336  0.4344598 -0.78478972
```

===>

-On 6 classe alors 6 moyennes on remarque que les classes 4 et 5 et 6 contiennent la majorité des moyennes avec des valeurs négatives ce qui indique que ces classes regroupent les valeurs faibles

- les moyennes de la classe 1 presque toutes positives Puis regroupent les variables élevé.

3.2.2 le taux d'inertie avec 6 classe :

$$between_{s/total_{ss}} = 50.6\% \quad (1)$$

==>le taux d'inertie >50% la plupart d'inertie totale est explique.

-Le taux d'inertie augmente avec l'augmentation de nombre de classe

3.2.3 L'enertie expliqué

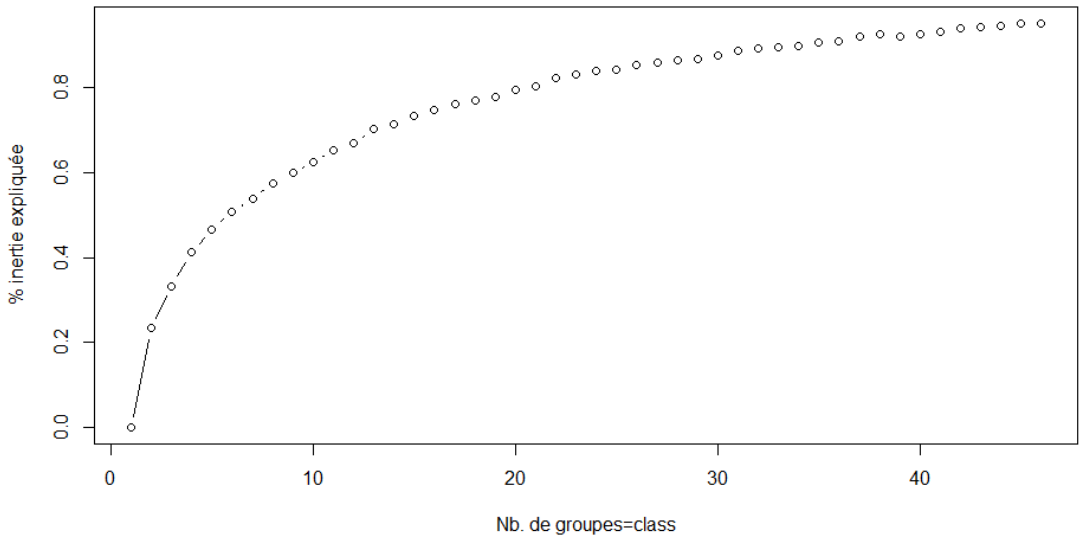
0.0000000 0.2337813 0.3333622 0.4120560 0.4506834 0.5082593

3.2.4 le nombre de classe N Avec max(inertie.expl)>0.95

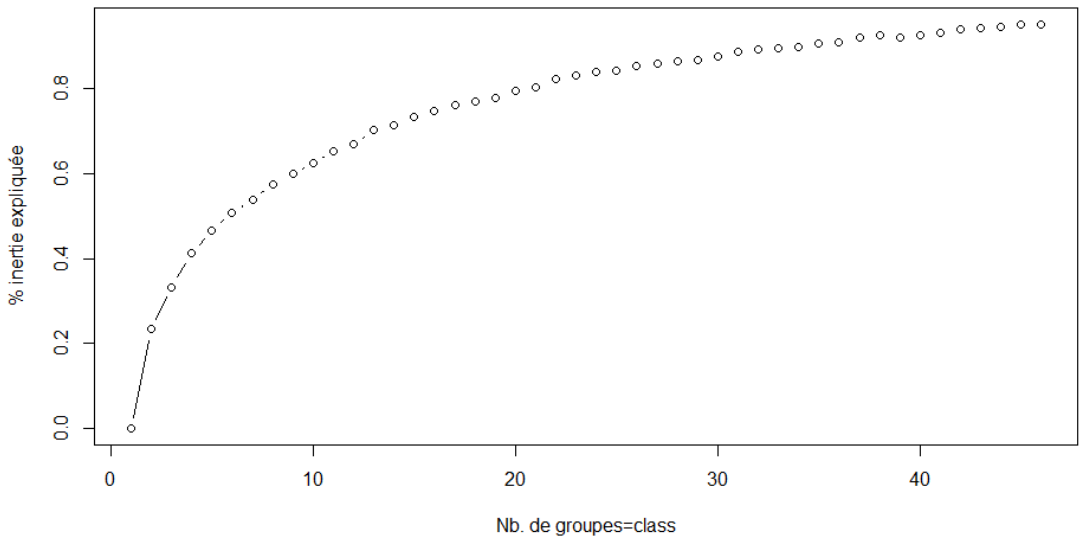
le critère de choisi N :est le plus petit entier tel que max(inertie.expl)>0.95

l'evolution de l'inertie explique en fonction de nombre de classes

-Avec **N=42** on a trouve **max(inertie.expl)=0.93**



-Avec **N=46** on a trouve **max(inertie.expl)=0.952>0.95**



==>le critère est vérifier

3.3 le nombre de classes avec le critère $\frac{\text{var}(I_2)}{\text{var}(I)} < 0,05$

le 2ème critère de choix du nombre de classes est le quotient de variance qui doit être<**5%**

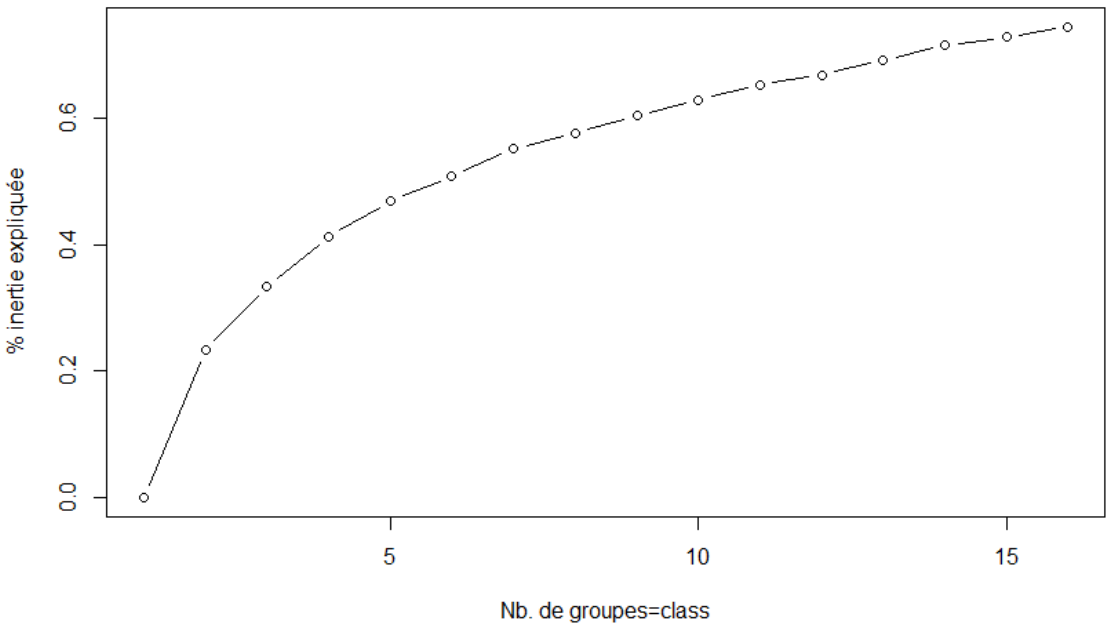
-Si on élimine apartir de **16** :

```
var(inertie.expl[16:N])*(N-16)*100/(var(inertie.expl)*(N-1))
  5.647233
```

-Si on élimine apartir de **17** :

```
var(inertie.expl[17:N])*(N-17)*100/(var(inertie.expl)*(N-1))
  4.77011
```

===> Alors nombre de classes retenus est **16**



3.4 La Classification Ascendante Hiérarchique CAH

-Principe de l'algorithme : construire dans chaque étape une partition de l'ensemble des individus en regroupant les éléments les plus proches.
1. On commence par calculer la dissimilarité entre les N objets.
2. Puis on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné, créant ainsi une classe comprenant ces deux objets.
3. On calcule ensuite la dissimilarité entre cette classe et les N − 2 autres objets en utilisant le critère d'agrégation. Puis on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation.

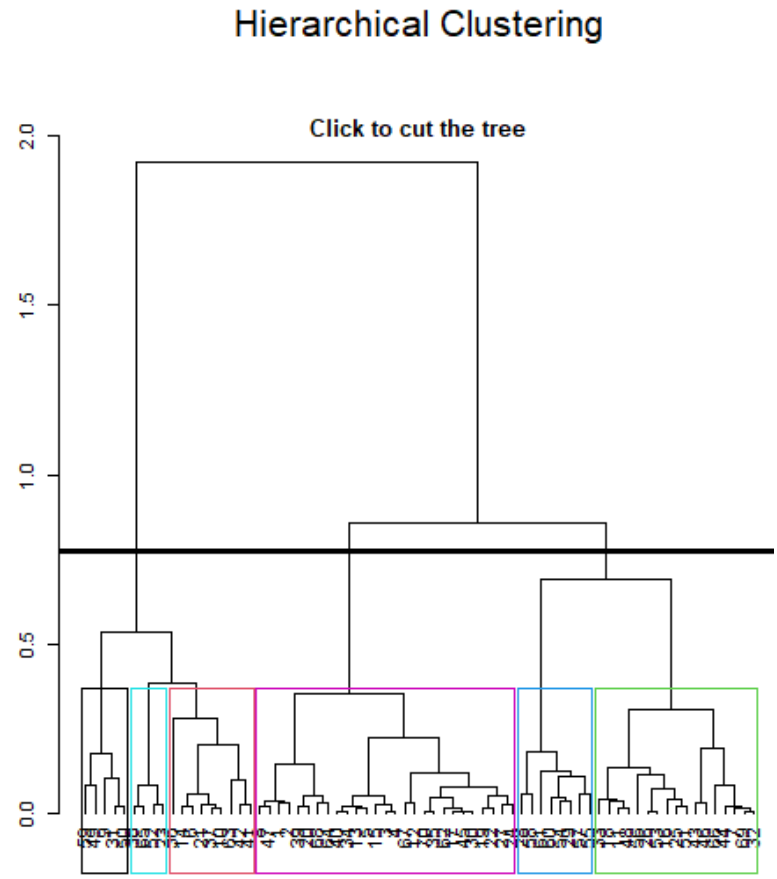
3.4.1 centré et réduire les variables

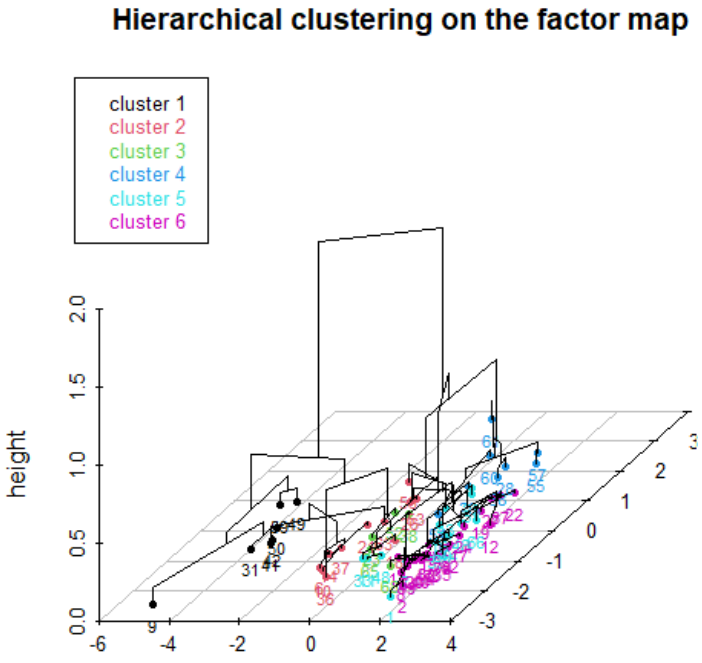
```
data.actifs_cr<-scale(data.actifs,center=T,scale=T)
```

```
> str(as.data.frame(data.actifs_cr))
'data.frame':   70 obs. of  9 variables:
 $ turda                                 : num  0.0644 -1.0086 -0.0502 0.0305 -1.7159 ...
 $ X2.house.age                          : num  1.1627 0.0662 -0.4777 -0.4777 -1.2058 ...
 $ X3.distance.to.the.nearest.MRT.station: num  -0.777 -0.613 -0.425 -0.425 -0.551 ...
 $ X4.number.of.convenience.stores       : num  2.141 1.768 0.272 0.272 0.272 ...
 $ X5.latitude                           : num  0.966 0.765 1.308 1.308 0.691 ...
 $ X6.longitude                          : num  0.436 0.391 0.662 0.662 0.572 ...
 $ dephm                                 : num  -0.861 -0.836 -0.83 -0.803 -0.8 ...
 $ SOLID                                 : num  -2.049 -1.3974 -0.0806 -0.0635 0.0701 ...
 $ Condi                                 : num  1.177 -0.109 -0.733 -1.678 -1.232 ...
```

3.4.2 critère du coupe (manuellement)

-1 première étape on coupe manuellement l'arbre on choisit les classes homogènes
-2 On coupe au niveau d'une longue branche
===>on a obtenue 6 classes :

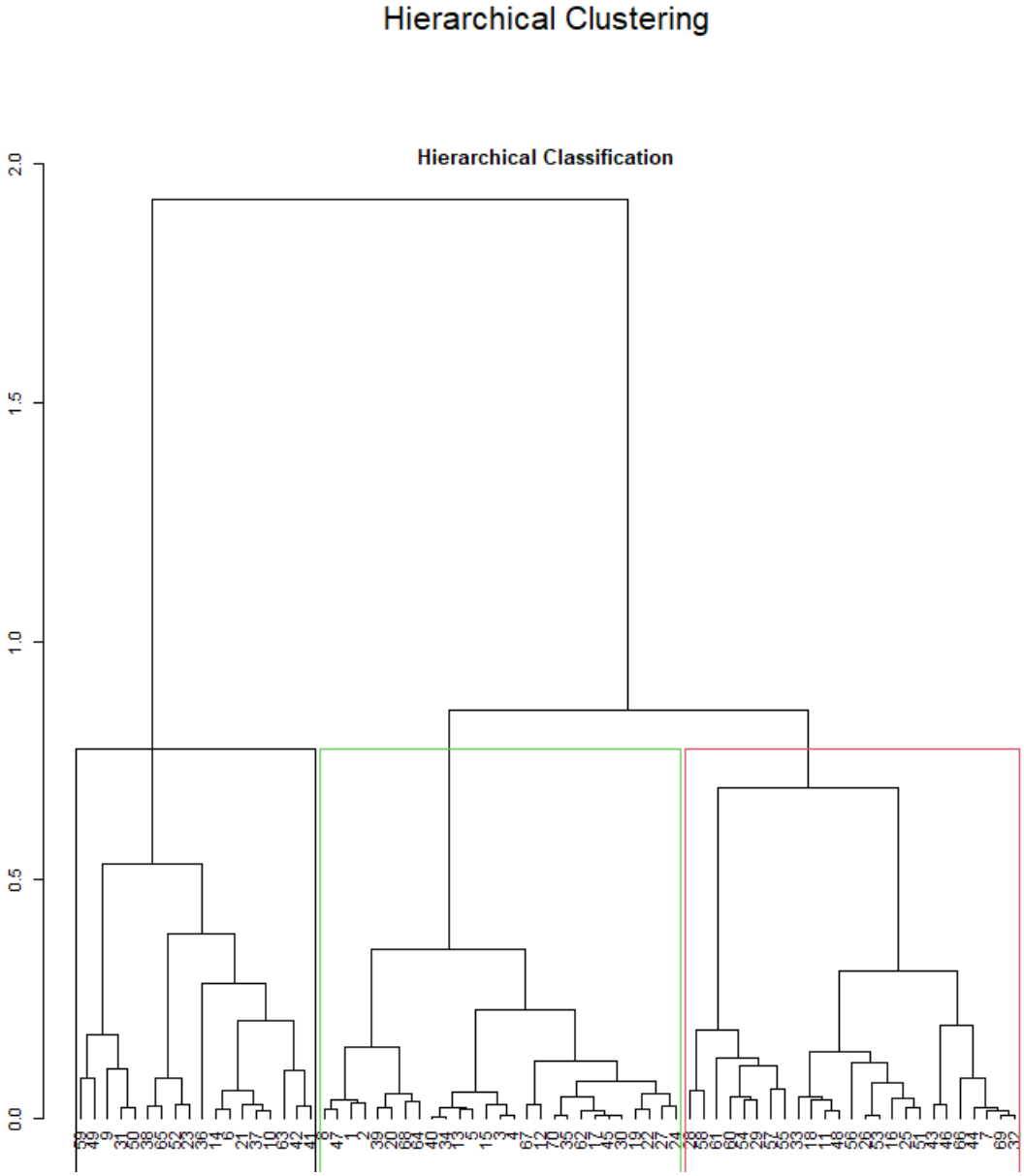

Hierarchical Clustering
Click to cut the tree
2.0
1.5
1.0
0.5
0.0

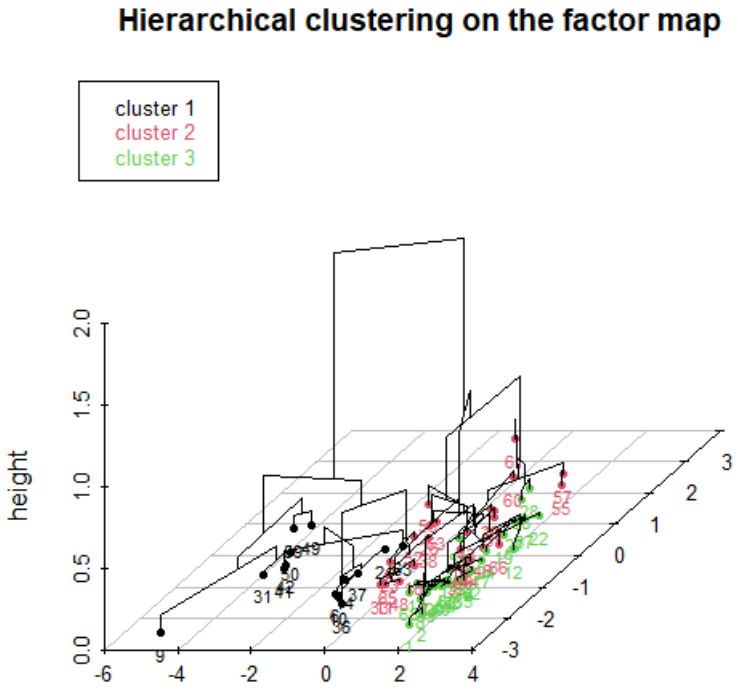

Hierarchical clustering on the factor map
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
height
2.0
1.5
1.0
0.5
0.0
-6
-4
-2
0
2
4
3
2
1
0
-1
-2
-3

3.4.3 Critère du coupe de François Husson

Res <- HCPC(as.data.frame(data.actifs),nb.clust =-1)

-on obtient 3 classes avec cettte coupe :





3.4.4 les variables quantitatives les plus corrélées avec la variable classification

Analyse de variance :

```
Link between the cluster variable and the quantitative variables
================================================================
                                            Eta2      P-value
X6.longitude                          0.78287334 6.024082e-23
X3.distance.to.the.nearest.MRT.station 0.77573734 1.779749e-22
X2.house.age                          0.46841279 6.408354e-10
X4.number.of.convenience.stores       0.44094642 3.464894e-09
X5.latitude                           0.21573991 2.913492e-04
turda                                 0.11550008 1.638235e-02
dephm                                 0.09284964 3.821694e-02
```

-on remarque que seules les variables quantitatives qui présentent une analyse de variance ont un facteur significatif par rapport à la variable de classification

==>les varaibles **Solid** et **Condi** ne présent pas car ne sont pas significatif par l'analyse de variance avec le variable quantitative de classification

Corrélation

D'après le tableau précicdent ona la corelation **Etat2** :

	Eta2
X6. longitude	0.78287334
X3. distance. to. the. nearest. MRT. station	0.77573734
X2. house. age	0.46841279
X4. number. of. conveni ence. stores	0.44094642
X5. 7atitude	0.21573991
turda	0.11550008
dephm	0.09284964

==>donc les varaiabes **X6.longitude** et **X3. distance. to. the. nearest.MRT.station** les plus corrélées avec la variable classification

3.4.5 la description des classes retenues par la varaiable qualitatives

```
Description of each cluster by quantitative variables
======================================================
$`1`
                                        v.test Mean in category  Overall mean sd in category Overall sd      p.value
X3.distance.to.the.nearest.MRT.station  7.229963       1.7282917 -4.594539e-17      0.8998377  0.9928314 4.831258e-13
X5.latitude                            -3.626551      -0.8669114 -4.645003e-14      1.3390329  0.9928314 2.872325e-04
X4.number.of.convenience.stores        -4.333640      -1.0359382  2.220446e-17      0.5238808  0.9928314 1.466637e-05
X6.longitude                           -7.251216      -1.7333722  3.990956e-13      0.5967121  0.9928314 4.130464e-13


$`2`
              v.test Mean in category  Overall mean sd in category Overall sd      p.value
X2.house.age 5.194946       0.8077402  8.624054e-18      0.8188269  0.9928314 2.047794e-07
turda        2.800634       0.4354587 -8.964055e-14      0.8482860  0.9928314 5.100238e-03
dephm        2.531099       0.3935499  3.682570e-17      1.1256129  0.9928314 1.137057e-02


$`3`
                                        v.test Mean in category  Overall mean sd in category Overall sd      p.value
X4.number.of.convenience.stores         4.952469       0.6835056  2.220446e-17      0.5898340  0.9928314 7.327785e-07
X6.longitude                            4.220830       0.5825299  3.990956e-13      0.2028792  0.9928314 2.434041e-05
X5.latitude                             2.757366       0.3805527 -4.645003e-14      0.6570157  0.9928314 5.826910e-03
turda                                  -2.129265      -0.2938665 -8.964055e-14      0.8230847  0.9928314 3.323236e-02
X3.distance.to.the.nearest.MRT.station -4.139863      -0.5713553 -4.594539e-17      0.1624278  0.9928314 3.475137e-05
X2.house.age                           -5.181605      -0.7151294  8.624054e-18      0.6391993  0.9928314 2.199853e-07
```

-la classe 1 regroupe les batiments qui ont une faible dimention(distance,longitude...)
-la classe 3 regroupe les bâtiments à grande dimension

3.4.6 calcule des taux d'inertie avant et après la consolidation de la CAH.

-la varaiance c'esi au niveau du inertie inter alors :

	Avant consolidation de la CAH.	après consolidation de la CAH.
Inertie inter	2.780738	2.928891

===> il y a une amélioration de l'inter inertie, elle est passer de 2,78 à 2,92 après consolidation des k-means

3.4.7 Comparaison Kmeans VS CAH

-Le clustering K-Means nécessite une connaissance préalable de K, c'est-à-dire du nombre de clusters que l'on veut diviser dans les données

-Dans le clustering hiérarchique CAH, on peut s'arrêter à n'importe quel nombre de clusters, que l'on trouve approprié en interprétant le dendrogramme.

-Dans notre cas avec Kmeans on a trouver N=46 classe (un nombre qui est trés grand),et avec CAH le nombre de class retenue est N=16