ALZHEIMER FEATURES PREDICTION

Takwa BOUCHADDEKH
Souleima HAWEL
Manar ELADEL
Rouaida SAHLI

2eme INFO_B

INTRODUCTION :

La maladie d'**Alzheimer** est une maladie neurodégénérative progressive qui affecte les fonctions cognitives, la mémoire et le comportement. La détection précoce de la maladie d'Alzheimer est cruciale pour une intervention rapide et de meilleurs résultats pour les patients. Les modèles d'apprentissage automatique peuvent jouer un rôle important dans la prévision de la probabilité de maladie d'Alzheimer sur la base de caractéristiques pertinentes.

Ce projet présente une analyse des caractéristiques liées à la maladie d'Alzheimer à l'aide d'un ensemble de données comprenant 373 échantillons et 10 colonnes. L'objectif principal est de prédire la variable cible « Groupe », qui classe les individus en déments, non déments ou convertis. L'ensemble de données comprend divers attributs démographiques et cliniques tels que le sexe, l'âge, l'éducation, le statut socio-économique et les scores d'évaluation cognitive. Les données englobent des variables catégorielles nominales, discrètes et continues, fournissant un ensemble diversifié d'informations pour l'analyse. Le code présenté sert de cadre fondamental pour l'analyse de la maladie d'Alzheimer, démontrant l'utilisation de techniques d'apprentissage automatique pour tirer des informations de l'ensemble de données fourni. Une personnalisation et un raffinement supplémentaires peuvent être appliqués en fonction d'objectifs de recherche spécifiques et des caractéristiques de l'ensemble de données.

FEATURES :

L'ensemble de données contient des informations sur 373 individus et comprend les colonnes suivantes :

Groupe : Variable cible indiquant le groupe de diagnostic (Dément, Non dément, Converti).

H/F : Sexe des individus (Homme/Femme).

Age : Âge des individus.

EDUC : Années d'études.

SES : statut socioéconomique, allant de 1 (faible) à 5 (élevé).

MMSE : Mini examen de l'état mental.

CDR : évaluation clinique de la démence.

eTIV : volume intracrânien total estimé.

nWBV : Volume du cerveau entier normalisé.

ASF : Facteur d’échelle Atlas.

Source: https://www.kaggle.com/datasets/brsdincer/alzheimer-features

I. ACP ET CLASSIFICATION :

Les statistiques avec python Manipulation des données avec Pandas :

Group	M/F	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
Nondemented	M	87	14	2.0	27.0	0.0	1987	0.696	0.883
Nondemented	M	88	14	2.0	30.0	0.0	2004	0.681	0.876
Demented	M	75	12	NaN	23.0	0.5	1678	0.736	1.046
Demented	M	76	12	NaN	28.0	0.5	1738	0.713	NaN
Demented	M	80	12	NaN	22.0	0.5	1698	0.701	1.034
...	...	...	...	...	...	...	...	...	...
Demented	M	82	16	1.0	28.0	0.5	1693	0.694	1.037
Demented	M	86	16	1.0	26.0	0.5	1688	0.675	2024-04-01 00:00:00
Nondemented	F	61	13	2.0	30.0	0.0	1319	0.801	1.331
Nondemented	F	63	13	2.0	30.0	0.0	1327	0.796	1.323
Nondemented	F	65	13	2.0	30.0	0.0	1333	0.801	1.317

On a remarqué l'absence de certaines valeurs on les a calculé leur pourcentage

Pourcentage des valeurs manquants= 0.655347036044087% < 1%

Donc la meilleure pratique est de **les supprimer**.

Dimension , Description et Info :

```
(354, 9)
```

	M/F	Age	EDUC	SES	MMSE	CDR
count	354	354.000000	354.000000	354.000000	354.000000	354.000000
unique	2	NaN	NaN	NaN	NaN	NaN
top	F	NaN	NaN	NaN	NaN	NaN
freq	204	NaN	NaN	NaN	NaN	NaN
mean	NaN	77.033898	14.703390	2.460452	27.409605	0.271186
std	NaN	7.811808	2.895662	1.134005	3.712626	0.370537
min	NaN	60.000000	6.000000	1.000000	4.000000	0.000000
25%	NaN	71.000000	12.000000	2.000000	27.000000	0.000000
50%	NaN	77.000000	15.000000	2.000000	29.000000	0.000000
75%	NaN	82.000000	16.750000	3.000000	30.000000	0.500000
max	NaN	98.000000	23.000000	5.000000	30.000000	2.000000

	eTIV	nWBV	ASF
count	354.000000	354.000000	354
unique	NaN	NaN	255
top	NaN	NaN	1.184
freq	NaN	NaN	5
mean	1489.991525	0.729879	NaN
std	175.768462	0.037842	NaN
min	1106.000000	0.644000	NaN
25%	1358.250000	0.699000	NaN
50%	1470.500000	0.729000	NaN
75%	1595.250000	0.757000	NaN
max	2004.000000	0.837000	NaN

```
Data columns (total 9 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   M/F     373 non-null    object
 1   Age     373 non-null    int64
 2   EDUC    373 non-null    int64
 3   SES     354 non-null    float64
 4   MMSE    371 non-null    float64
 5   CDR     373 non-null    float64
 6   eTIV    373 non-null    int64
 7   nWBV    373 non-null    float64
 8   ASF     372 non-null    object
dtypes: float64(4), int64(3), object(2)
```

la matrice de corrélation entre les colonnes numériques

```
           Age      EDUC       SES      MMSE       CDR      eTIV      nWBV  \
Age   1.000000 -0.028428 -0.029778  0.057317 -0.020790  0.028970 -0.517986
EDUC -0.028428  1.000000 -0.722100  0.175408 -0.116189  0.262189 -0.023380
SES  -0.029778 -0.722100  1.000000 -0.141313  0.062204 -0.247971  0.080364
MMSE  0.057317  0.175408 -0.141313  1.000000 -0.705719 -0.025977  0.342763
CDR  -0.020790 -0.116189  0.062204 -0.705719  1.000000  0.059386 -0.352649
eTIV  0.028970  0.262189 -0.247971 -0.025977  0.059386  1.000000 -0.199442
nWBV -0.517986 -0.023380  0.080364  0.342763 -0.352649 -0.199442  1.000000
ASF  -0.020646 -0.246135  0.242839  0.034114 -0.069958 -0.988838  0.203859

           ASF
Age  -0.020646
EDUC -0.246135
SES   0.242839
MMSE  0.034114
CDR  -0.069958
eTIV -0.988838
nWBV  0.203859
ASF   1.000000
```

Les caractéristiques les plus corrélées :

eTIV et ASF : Ces caractéristiques présentent une corrélation négative forte de **-0.989**, ce qui indique une relation inverse entre le volume intracrânien estimé (eTIV) et le facteur de mise à l'échelle du cerveau (ASF). Cela signifie que des valeurs élevées d'eTIV sont associées à des valeurs faibles d'ASF et vice versa.

Les caractéristiques les moins corrélées :

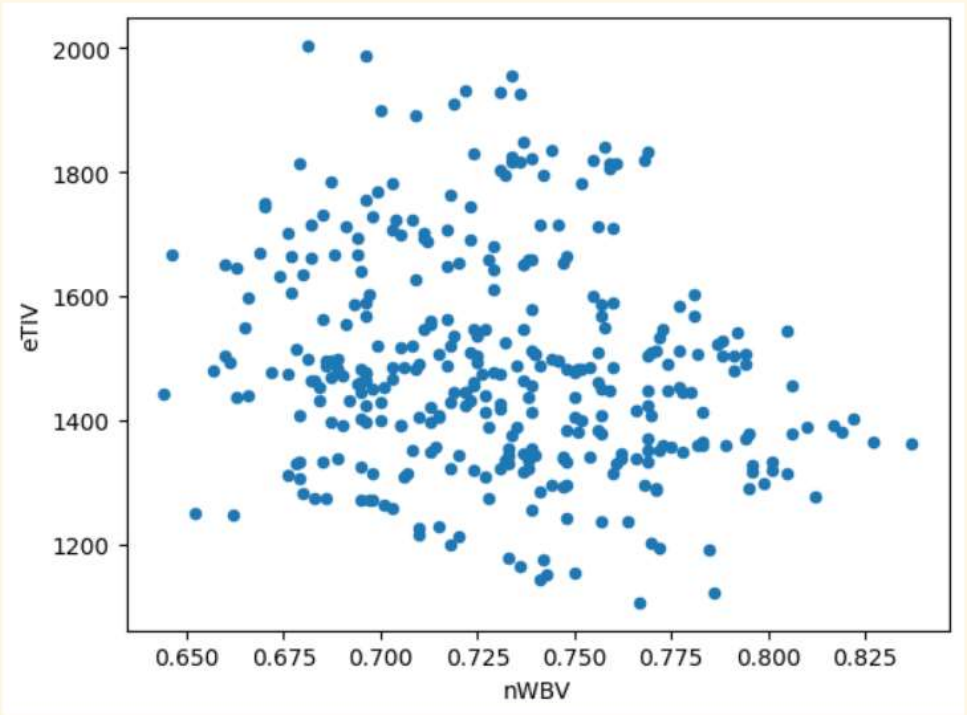
Age et ASF : La corrélation entre ces caractéristiques est également très faible à **-0.021**, ce qui indique qu'il n'y a pas de relation linéaire forte entre l'âge et le facteur de mise à l'échelle du cerveau (ASF).

transformation – centrage-réduction :

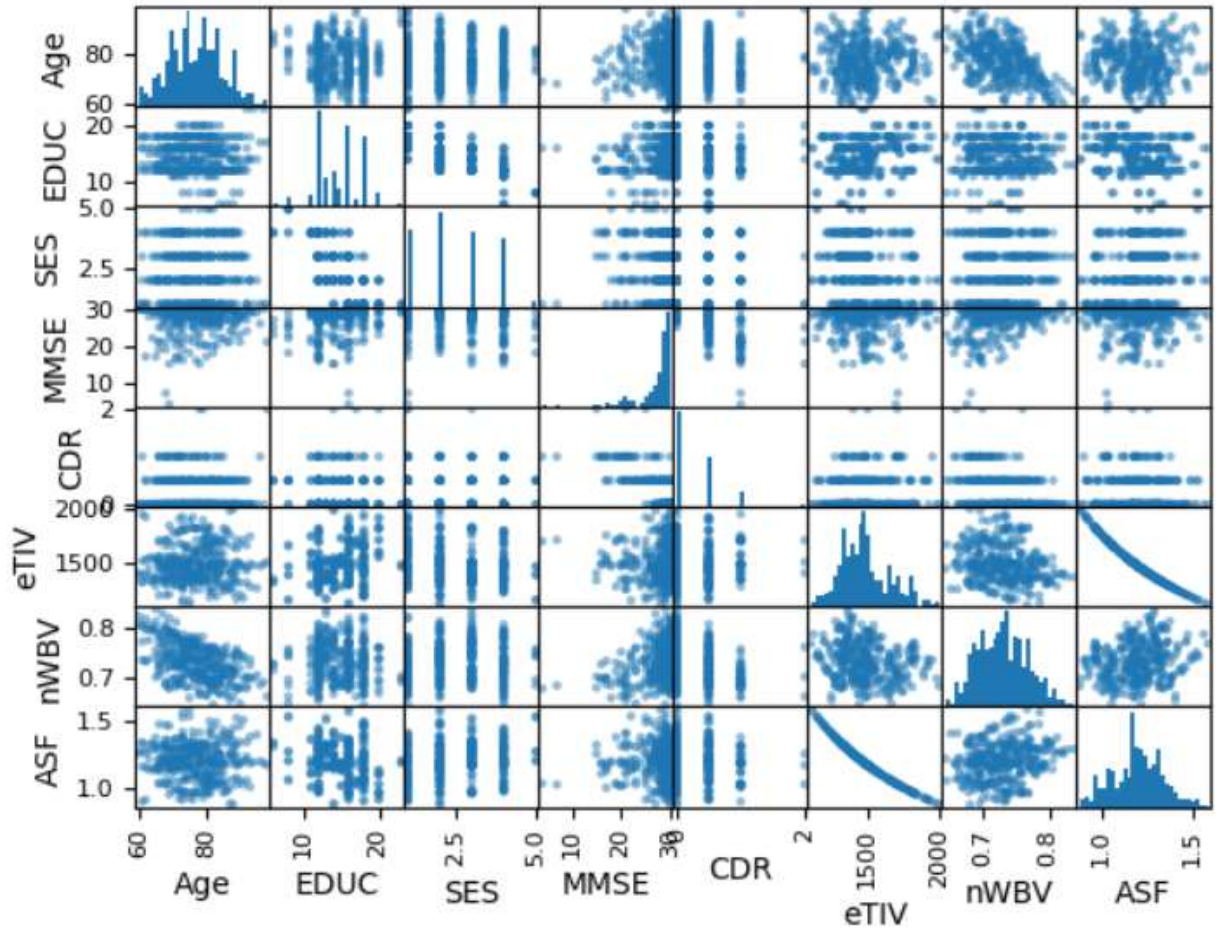
```
[[ 1.28499166 -0.23876315 -0.41956694 ...  2.84806485 -0.90808279
  -2.29191196]
 [ 1.41312051 -0.23876315 -0.41956694 ...  2.94486213 -1.30486252
  -2.3431111 ]
 [ 1.41312051  1.14386528  0.47104754 ... -1.54767029 -0.53775504
   1.81133341]
 ...
 [-2.04635847 -0.58442025 -0.41956694 ... -0.95549872  1.86937534
   0.98483301]
 [-1.79010077 -0.58442025 -0.41956694 ... -0.90994706  1.73711543
   0.9263197 ]
 [-1.53384307 -0.58442025 -0.41956694 ... -0.87578332  1.86937534
   0.88243473]]
```

Nuage de points :

```
<Axes: xlabel='nWBV', ylabel='eTIV'>
```



Scatter_matrix :



Les diagrammes de dispersion dans la matrice peut confirmer les relations entre les variables qui sont cohérentes avec ce que nous avons observé dans la matrice de corrélation.

eTIV et ASF : Étant donné que ces deux caractéristiques sont fortement négativement corrélées, vous pouvez vous attendre à voir un schéma où les points se regroupent également de manière inversement proportionnelle. À mesure que le volume intracrânien estimé (eTIV) augmente, vous vous attendez à voir le facteur de mise à l'échelle du cerveau (ASF) diminuer et vice versa.

Age et ASF : Étant donné que la corrélation entre ces deux caractéristiques est très faible (-0.021), le nuage de points montre une dispersion aléatoire des points

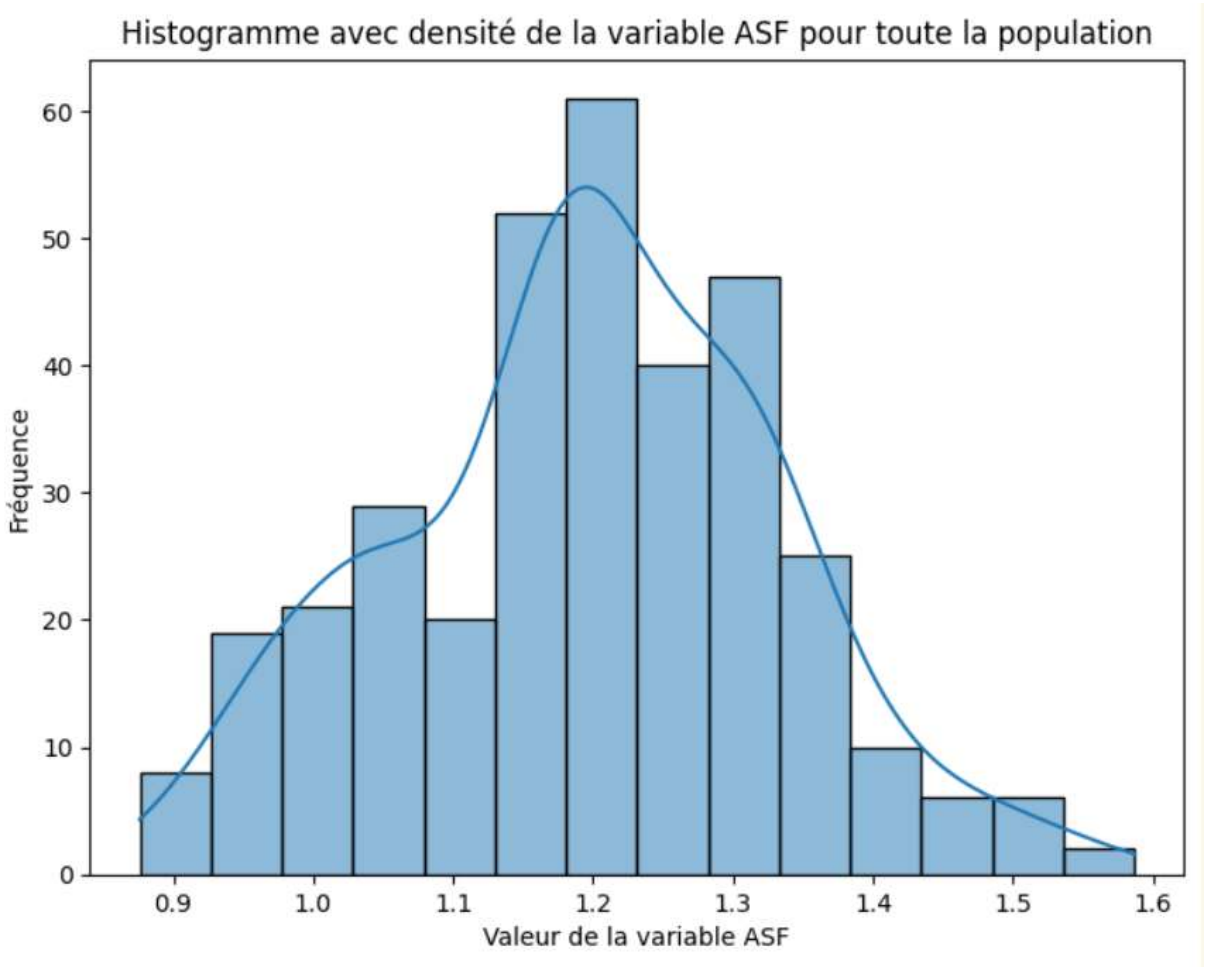
Statistiques selon l'espèce MMSE :

	MMSE							
	count	mean	std	min	25%	50%	75%	max
Group								
Converted	37.0	28.675676	1.564432	24.0	28.0	29.0	30.0	30.0
Demented	127.0	24.322835	4.657954	4.0	21.0	26.0	28.0	30.0
Nondemented	190.0	29.226316	0.882722	26.0	29.0	29.0	30.0	30.0

Ces statistiques résument les scores MMSE (Mini-Mental State Examination) en fonction de la catégorie de l'espèce :

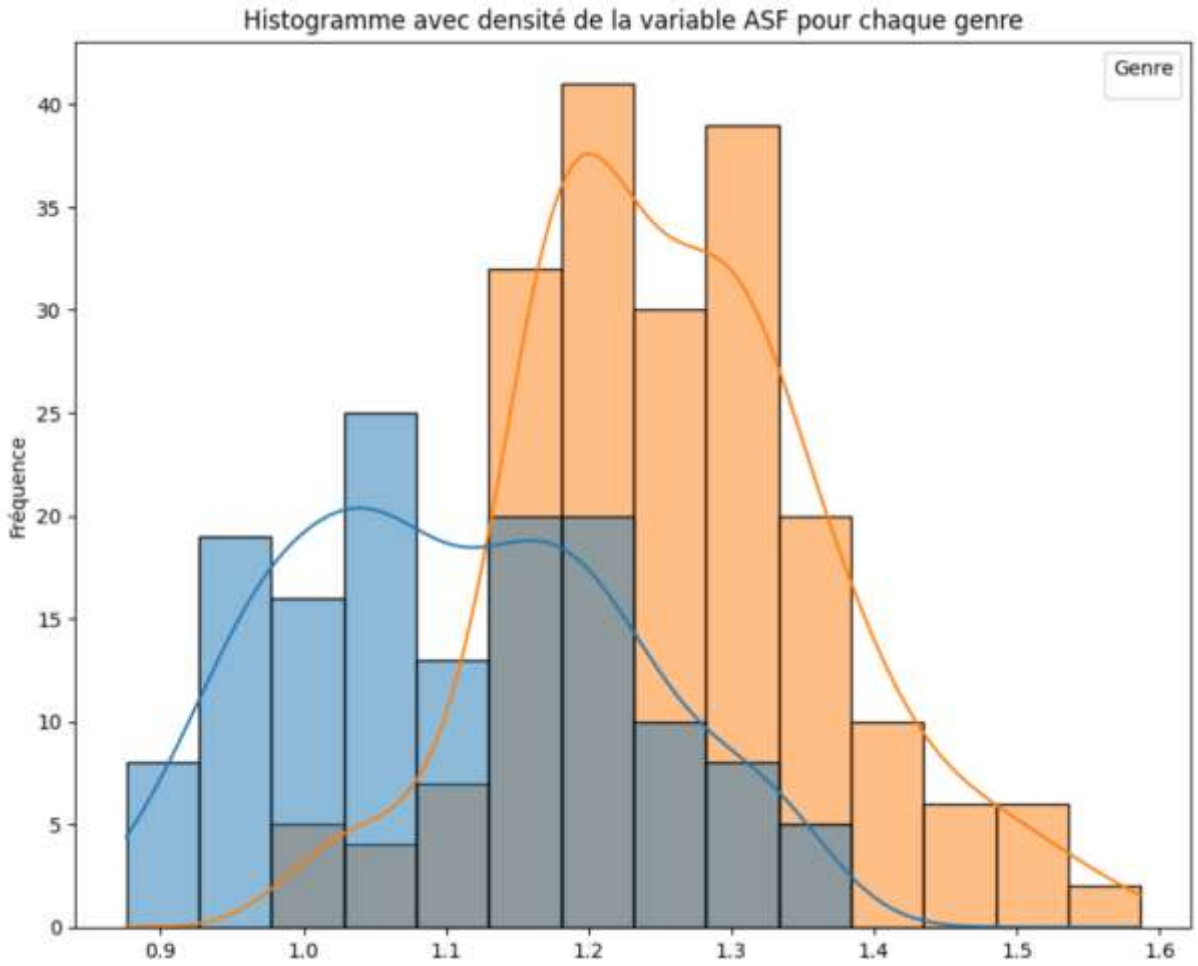
1. **Converted** : Il y a 37 individus dans cette catégorie. En moyenne, leur score MMSE est d'environ 28.68, avec un écart-type de 1.56. Les scores varient de 24 à 30, avec la plupart des scores situés entre 28 et 30.
2. **Demented** : Cette catégorie comprend 127 individus. Leur score MMSE moyen est d'environ 24.32, avec un écart-type de 4.66. Les scores MMSE varient considérablement, allant de 4 à 30, mais la plupart se situent entre 21 et 28.
3. **Nondemented** : Il y a 190 individus dans cette catégorie. En moyenne, leur score MMSE est d'environ 29.23, avec un écart-type de 0.88. La plupart des scores MMSE se situent entre 29 et 30, avec quelques scores plus bas à 26.

Histogramme avec densité de la variable ‘ASF’ pour toute la population :



L'histogramme pour toute la population montre que la variable ASF est distribuée de manière approximativement normale, avec une légère asymétrie vers la droite. Cela signifie que la plupart des valeurs de l'ASF se situent autour de la moyenne, avec un nombre plus petit de valeurs extrêmes. La densité de la distribution confirme cette observation, avec un pic central et des queues décroissantes vers les valeurs extrêmes.

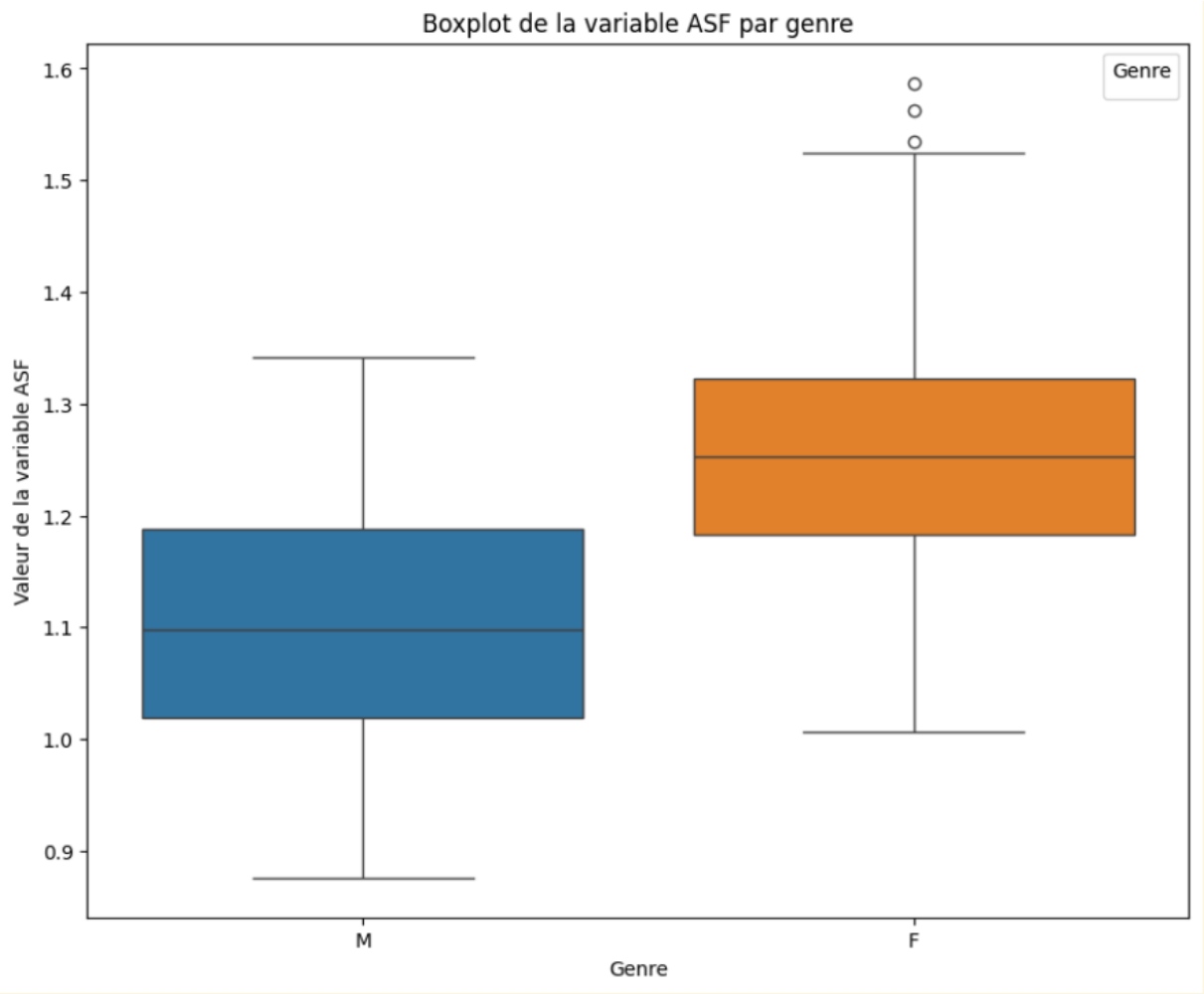
Histogramme avec densité de la variable ASF pour chaque genre :



On observe une légère différence de distribution entre les genres:

- La distribution pour les hommes semble être légèrement plus étalée que celle pour les femmes, ce qui suggère une plus grande variabilité de l'ASF chez les hommes.
- La densité de la distribution confirme également cette observation, avec des pics légèrement plus larges pour les hommes

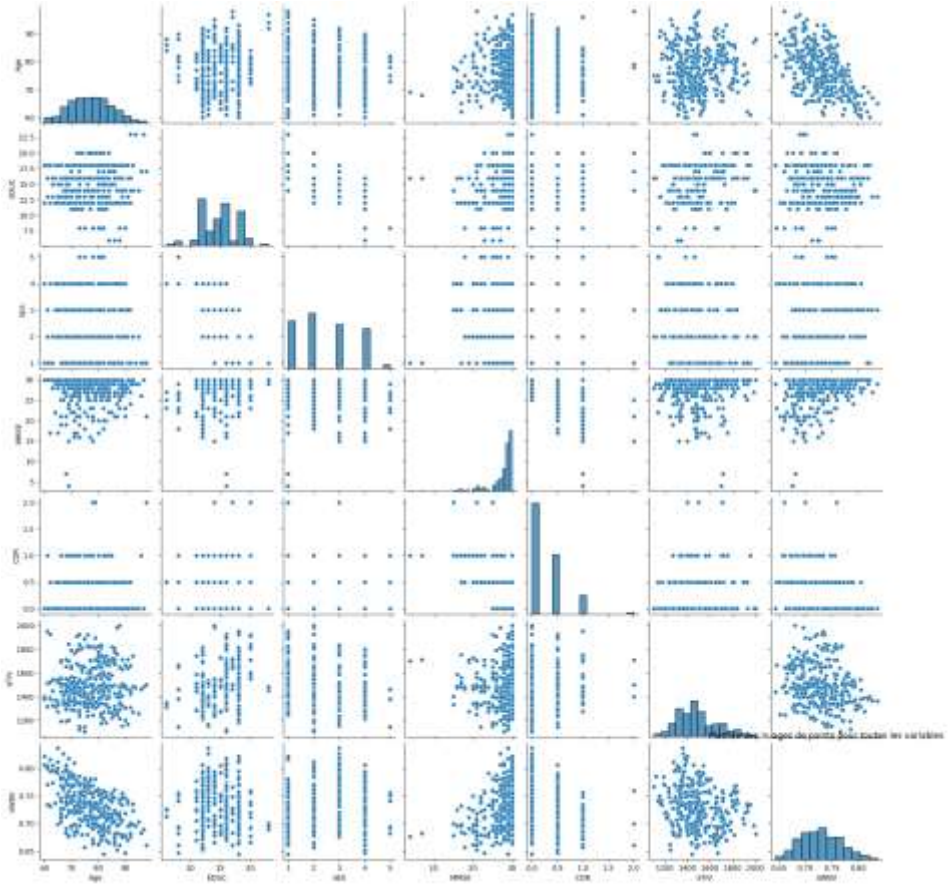
Comparaison des distributions avec un boxplot :



Ce boxplot compare la répartition de la variable ASF entre hommes et femmes :

Les valeurs médianes de l'ASF sont légèrement plus élevées chez les hommes (environ 1,25) que chez les femmes (environ 1,15).

Matrice des nuages de points pour toutes les variables :



Choix des axes factoriels :

Après avoir Calculer la matrice de corrélation et les valeurs et vecteurs propres et le pourcentage de variance expliquée par chacun des axes factoriels on obtient :

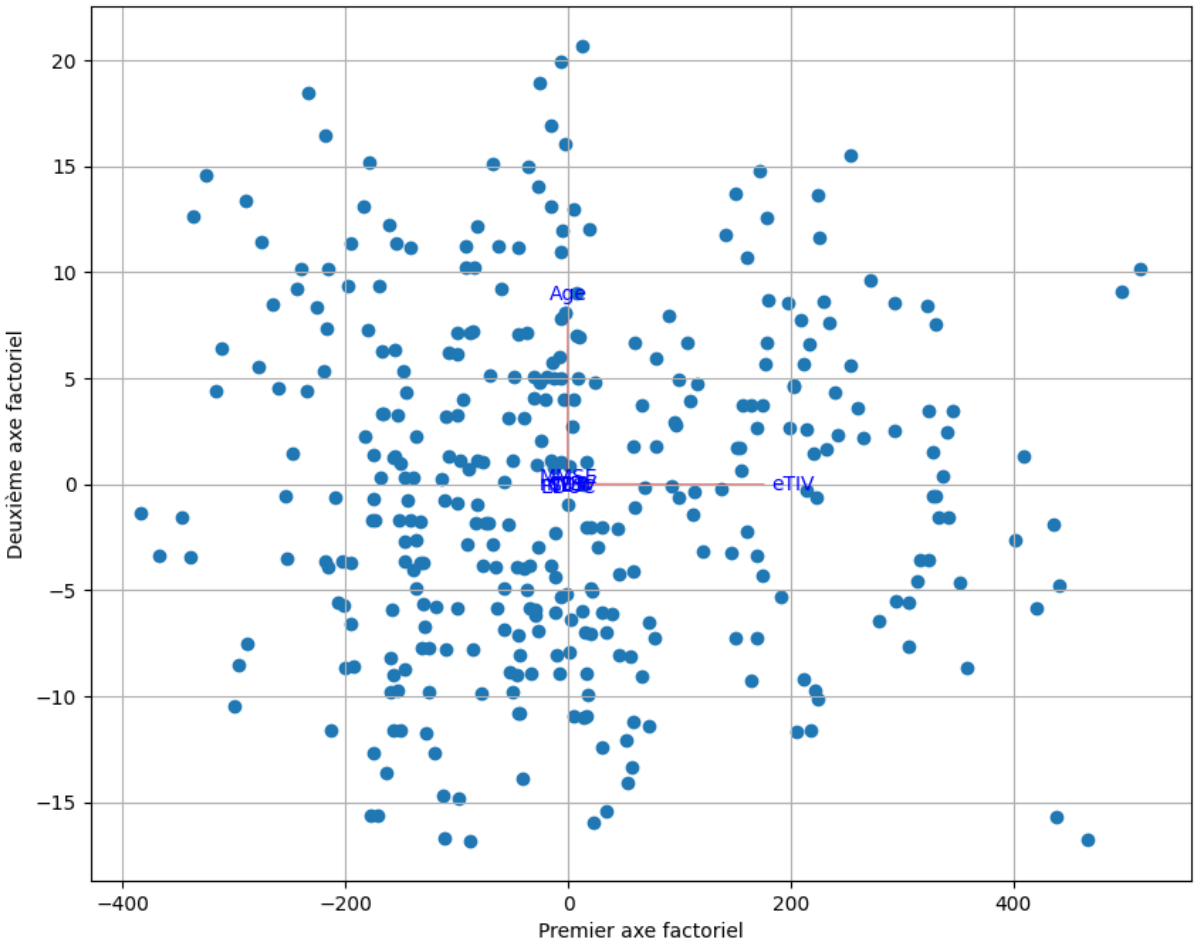
	valprop	inertie	inertiecum
0	2.105095	30.072783	30.072783
1	1.906757	27.239391	57.312174
2	1.282790	18.325575	75.637749
3	0.807652	11.537881	87.175630
4	0.340680	4.866862	92.042492
5	0.287551	4.107870	96.150362
6	0.269475	3.849638	100.000000

Les valeurs propres dans une analyse en composantes principales (ACP) représentent la quantité de variance expliquée par chaque composante principale (axe factoriel).Plus la valeur propre est élevée, plus l'axe correspondant capture une plus grande part de la variance totale des données. Les valeurs propres sont ordonnées de manière décroissante : la première valeur propre représente la plus grande quantité de variance expliquée, suivie de la deuxième, et ainsi de suite.

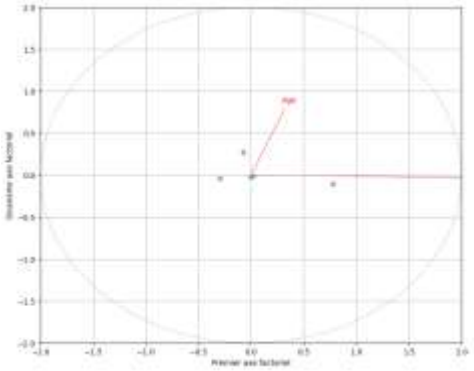
Pour calculer le pourcentage de variance expliquée par chaque axe factoriel, vous pouvez diviser chaque valeur propre par la somme totale des valeurs propres, puis multiplier par 100 pour obtenir le pourcentage.

En regardant ces pourcentages, nous pouvons déduire que le nombre de **dimensions du sous-espace vectoriel de projection** devrait être choisi de manière à expliquer un pourcentage significatif de la variance totale. Dans ce cas, une **dimension de 3 ou 4** pourrait être choisie, car cela expliquerait environ **75% à 87% de la variance totale des données**. Ce résultat était prévisible car il est souvent recommandé de sélectionner un nombre suffisant de composantes principales pour expliquer une proportion significative de la variance totale, tout en évitant la surdimensionnalité.

Projections des individus sur le plan correspondant aux deux premiers axes factoriels :



le graphique avec le cercle unité



Le graphique montre comment les individus sont positionnés par rapport à deux axes principaux qui résument les relations entre les variables.

les individus regroupés ensemble partagent probablement des caractéristiques similaires, tandis que ceux qui sont éloignés les uns des autres peuvent différer davantage.

Le premier axe semble être influencé par certaines variables, comme ‘Age’ , tandis que le deuxième axe semble être influencé principalement par ‘EDUC’.

La méthode de classification K-means avec 3 groupes (clusters) :

```
     Age  EDUC  SES  MMSE  CDR    eTIV   nWBV
0   87.0  14.0  2.0  27.0  0.0  1987.0  0.696
0   81.0  11.0  4.0  28.0  0.0  1750.0  0.670
0   87.0  12.0  4.0  30.0  0.0  1762.0  0.718
0   86.0  12.0  4.0  29.0  0.0  1783.0  0.703
0   77.0  16.0  2.0  30.0  0.0  1628.0  0.709
..   ...   ...  ...   ...  ...     ...    ...
2   82.0  16.0  3.0  29.0  0.0  1484.0  0.760
2   73.0  13.0  2.0  23.0  0.5  1536.0  0.725
2   75.0  13.0  2.0  28.0  0.5  1520.0  0.708
2   75.0  12.0  3.0  28.0  0.5  1407.0  0.770
2   98.0  17.0  1.0  21.0  2.0  1503.0  0.660

[354 rows x 7 columns]
[[239.74697266 677.5746057  499.88894261]
 [256.78319979 694.59162127 516.91452592]
 [532.49675237  95.30142925 272.48125012]
 ...
 [428.79128645  18.30025053 168.9762473 ]
 [420.7195102   22.20364977 160.83208182]
 [414.65345219  26.2514866  154.69267521]]
     Classe           ID      DistG1      DistG2      DistG3
0         1  Nondemented  239.746973  677.574606  499.888943
1         1  Nondemented  256.783200  694.591621  516.914526
2         2  Nondemented  532.496752   95.301429  272.481250
...
352       2  Nondemented  420.719510   22.203650  160.832082
353       2  Nondemented  414.653452   26.251487  154.692675
```

Classe : Cette colonne représente les classes attribuées par l'algorithme k-means à chaque observation. Dans notre cas, il y a trois classes, numérotées de 1 à 3.

ID : Cette colonne contient les identifiants des observations, qui peuvent être des individus ou des échantillons.

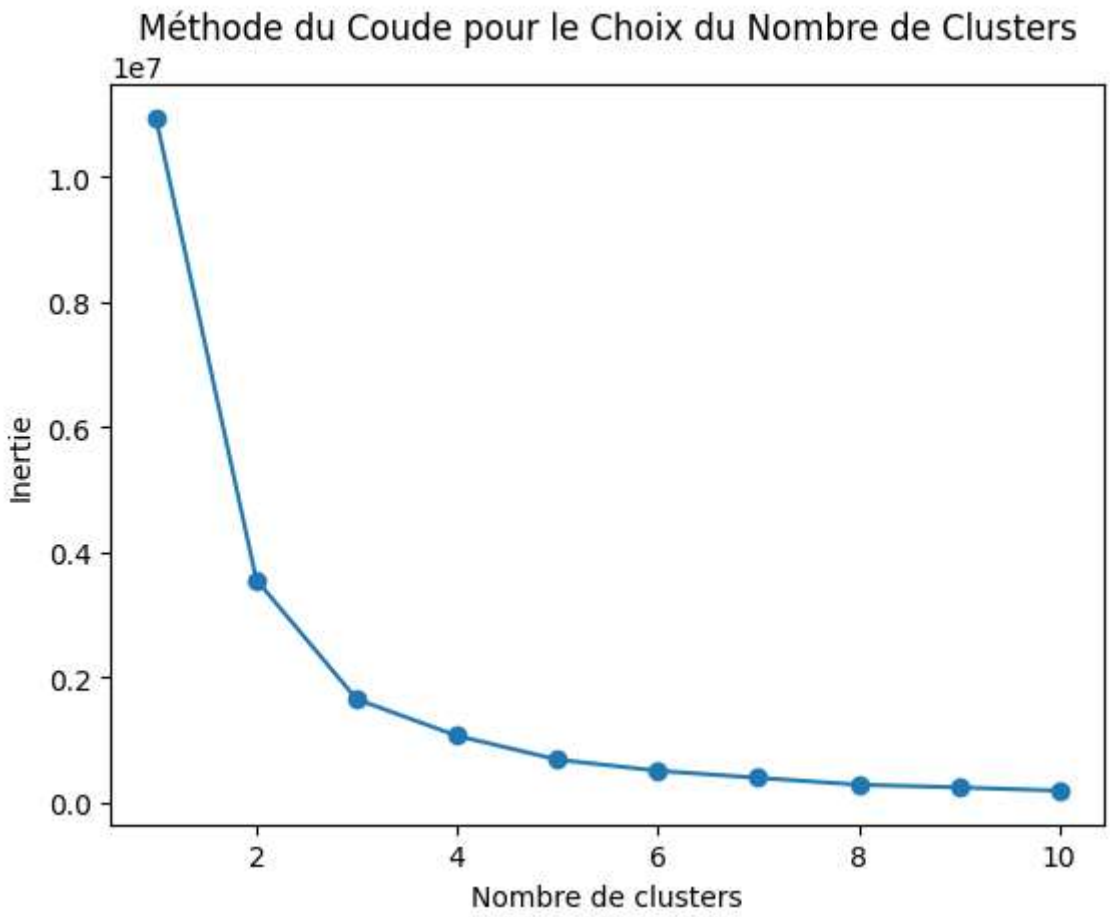
DistG1, **DistG2**, **DistG3** : Ces colonnes représentent les distances entre chaque observation et les centres des clusters (ou groupes) identifiés par l'algorithme k-means. Par exemple, **DistG1** représente la distance entre l'observation et le centre du premier cluster, **DistG2** représente la distance entre l'observation et le centre du deuxième cluster, et ainsi de suite.

La méthode de classification K-means avec 4 groupes (clusters) :

```
     Age  EDUC  SES  MMSE  CDR    eTIV   nWBV
0   87.0  14.0  2.0  27.0  0.0  1987.0  0.696
0   85.0  12.0  3.0  30.0  0.0  1820.0  0.755
0   86.0  12.0  3.0  27.0  0.0  1813.0  0.761
0   76.0  16.0  3.0  30.0  0.0  1832.0  0.769
0   72.0  20.0  1.0  26.0  0.5  1911.0  0.719
..   ...   ...  ...   ...  ...     ...    ...
3   77.0  20.0  1.0  23.0  1.0  1713.0  0.756
3   79.0  20.0  1.0  25.0  2.0  1710.0  0.760
3   80.0  20.0  1.0  29.0  0.0  1587.0  0.693
3   92.0  16.0  1.0  30.0  0.0  1662.0  0.682
3   66.0  16.0  1.0  19.0  1.0  1695.0  0.711

[354 rows x 7 columns]
     Classe           ID      DistG1      DistG2      DistG3      DistG4
0         0  Nondemented  141.477242  514.472788  685.444569  318.832474
1         0  Nondemented  158.495264  531.497523  702.461719  335.867091
2         2  Nondemented  631.044915  257.914322   87.542181  453.335053
3         2  Nondemented  646.083003  272.988915  102.662623  468.369753
4         3  Nondemented  157.044171  216.414726  387.383374   21.271193
..      ...          ...         ...         ...         ...         ...
349       3     Demented  153.041296  220.445514  391.411943   24.890068
350       3     Demented  158.249455  215.587170  386.497346   20.710605
351       2  Nondemented  527.178221  154.453216   23.209334  349.791179
352       2  Nondemented  519.125828  146.299198   28.698712  341.699248
353       2  Nondemented  513.079065  140.148679   33.393503  335.612724
```

Méthode du coude pour le choix du nombre de clusters :



Descente initiale: Au début, lorsque nous augmentons le nombre de clusters, l'inertie diminue rapidement. Cela signifie que les premiers clusters créés regroupent efficacement des points similaires.

Coude: Ensuite, nous observons un point où la baisse de l'inertie ralentit, formant ce que l'on appelle le "coude". Ce coude représente le nombre optimal de clusters où l'ajout de clusters supplémentaires ne donne plus d'amélioration significative.

Détermination du nombre optimal de clusters: En nous basant sur ce coude, nous pouvons estimer le nombre optimal de clusters. Dans ce cas, **le nombre optimal semble être 2**, car après ce point, l'amélioration de la structure des groupes est moins significative.

Interprétation des classes : statistiques comparatives

```
Moyennes des caractéristiques de chaque cluster :
               Age       EDUC       SES       MMSE       CDR         eTIV  \
cluster
0        76.666667  16.060606  2.181818  28.484848  0.196970  1843.181818
1        76.533333  14.466667  2.540000  27.166667  0.320000  1470.713333
2        76.198113  13.820755  2.811321  27.764151  0.198113  1301.641509
3        79.738462  16.000000  1.846154  26.846154  0.315385  1662.323077

              nWBV
cluster
0        0.728606
1        0.727793
2        0.744387
3        0.711677
```

```
Différences significatives entre les clusters :
               Age      EDUC       SES      MMSE       CDR        eTIV  \
cluster
1       -0.133333 -1.593939  0.358182 -1.318182  0.123030 -372.468485
2       -0.335220 -0.645912  0.271321  0.597484 -0.121887 -169.071824
3        3.540348  2.179245 -0.965167 -0.917997  0.117271  360.681567

              nWBV
cluster
1       -0.000813
2        0.016593
3       -0.032710

Caractéristiques les plus discriminantes pour chaque cluster :
Cluster 1:
Caractéristique la plus discriminante: SES, Différence: 0.3581818181818184
Cluster 2:
Caractéristique la plus discriminante: MMSE, Différence: 0.5974842767295598
Cluster 3:
Caractéristique la plus discriminante: eTIV, Différence: 360.68156748911474
```

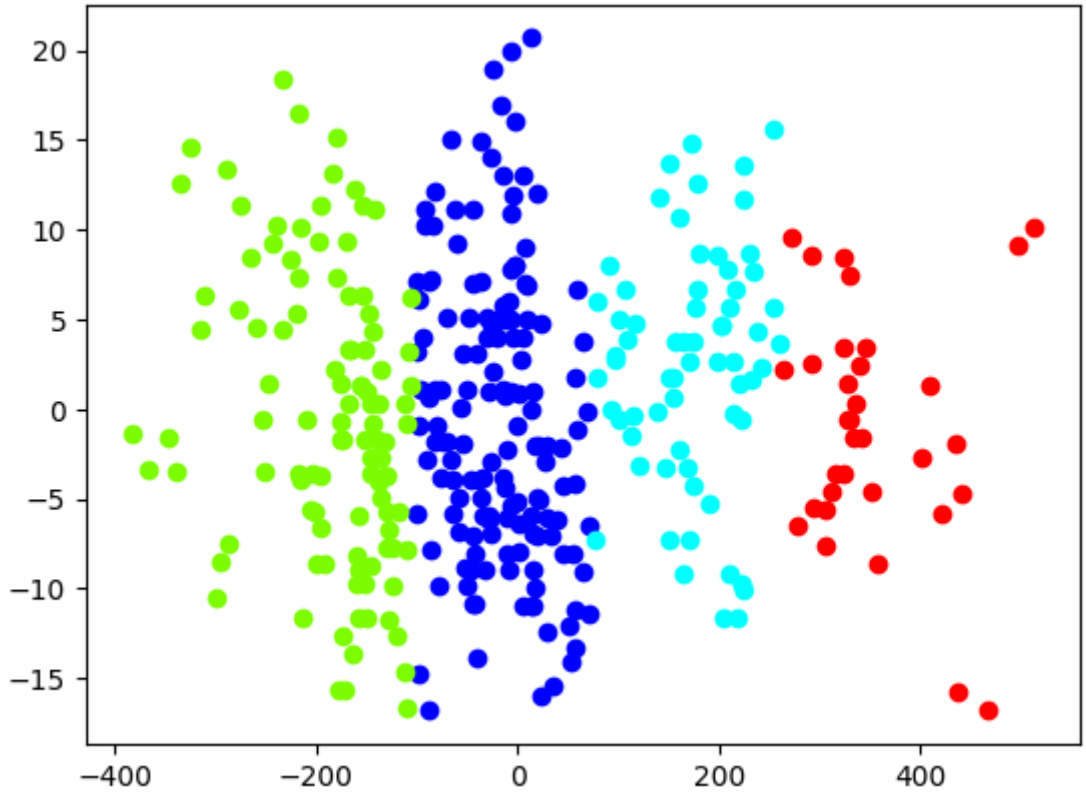
Cet algorithme K-means a été appliqué avec 4 clusters sur les données numériques. Les moyennes des caractéristiques de chaque cluster ont été calculées, mettant en évidence des différences significatives entre eux.

Le premier cluster se distingue par un niveau socio-économique (SES) relativement élevé.

Le deuxième cluster est caractérisé par un score MMSE (Mini Mental State Examination) plus faible.

Le troisième cluster se démarque par un volume intracrânien (eTIV) significativement plus élevé.

Projeter dans le plan factoriel :



Cette visualisation permet de voir la répartition des points dans l'espace des deux premières composantes principales, tout en mettant en évidence les regroupements de points qui représentent des clusters similaires.

II. AFC :

Pour une meilleure compréhension et interprétation des données relatives à la démence dans notre ensemble de données , Ce code Python utilise la fonction map_cdr_label pour mapper les valeurs numériques de la colonne 'CDR' à des étiquettes descriptives telles que 'moderate dementia', 'uncertain dementia', et 'No dementia'.

```
0      0.0
1      0.0
5      0.0
6      0.0
7      0.0
      ...
366    1.0
368    0.5
370    0.0
371    0.0
372    0.0
Name: CDR, Length: 346, dtype: float64
```

```
0              No dementia
1              No dementia
5              No dementia
6              No dementia
7              No dementia
                  ...
366     moderate dementia
368    uncertain dementia
370            No dementia
371            No dementia
372            No dementia
Name: CDR, Length: 346, dtype: object
```

Le tableau de contingence:

```
Tableau de contingence :
[[143  15  44]
 [ 62  19  63]]
```

	No dementia	moderate dementia	uncertain dementia
F	143	15	44
M	62	19	63

Cette interprétation nous donne un aperçu de la répartition des niveaux de démence selon le genre des individus dans notre ensemble de données.

le tableau de contingence en fréquence

```
Tableau de contingence en fréquence :
[[0.4132948  0.0433526  0.12716763]
 [0.17919075 0.05491329 0.18208092]]
```

Les pourcentages dans le tableau de contingence en fréquence représentent la proportion de chaque cellule par rapport au total de l'échantillon.

Ces pourcentages en tableau de contingence en fréquence permettent une comparaison directe des proportions de chaque catégorie de démence selon le genre.

Le tableau des profils-lignes :

```
Tableau des profils-lignes:
[[0.70792079 0.07425743 0.21782178]
 [0.43055556 0.13194444 0.4375    ]]
```

Les valeurs dans le tableau des profils-lignes sont les proportions relatives des différents niveaux de CDR pour chaque genre.

Dans la **première ligne** correspondant au **genre féminin ('F')** :

- Environ 70.79% des cas ont un CDR de "No dementia" (pas de démence)
- Environ 7.43% des cas ont un CDR de "moderate dementia" (démence modérée)
- Environ 21.78% des cas ont un CDR de "uncertain dementia" (démence incertaine)

Dans la **deuxième ligne** correspondant au **genre masculin ('M')** :

- Environ 43.06% des cas ont un CDR de "No dementia" (pas de démence)
- Environ 13.19% des cas ont un CDR de "moderate dementia" (démence modérée)
- Environ 43.75% des cas ont un CDR de "uncertain dementia" (démence incertaine)

Ces profils-lignes fournissent une vue détaillée des différences dans la distribution des niveaux de démence entre les genres féminin et masculin dans votre ensemble de données.

le tableau des profils-colonnes :

```
Tableau des profils-colonnes:
[[0.69756098 0.44117647 0.41121495]
 [0.30243902 0.55882353 0.58878505]]
```

Les valeurs dans le tableau des profils-colonnes sont les proportions relatives des différents genres pour chaque niveau de CDR.

Dans la première colonne correspondant au CDR "No dementia" :

- Environ 69.76% des individus ont un genre féminin ('F')
- Environ 30.24% des individus ont un genre masculin ('M')

Dans la deuxième colonne correspondant au CDR "moderate dementia" :

- Environ 44.12% des individus ont un genre féminin ('F')
- Environ 55.88% des individus ont un genre masculin ('M')

Dans la troisième colonne correspondant au CDR "uncertain dementia" :

- Environ 41.12% des individus ont un genre féminin ('F')
- Environ 58.79% des individus ont un genre masculin ('M')

Ces profils-colonnes fournissent une vue détaillée des différences dans la distribution des genres parmi les différents niveaux de démence dans votre ensemble de données.

Tableaux des fréquences théoriques :

```
Tableau des fréquences théoriques :
[[119.68208092  19.84971098  62.46820809]
 [ 85.31791908  14.15028902  44.53179191]]
```

Chaque cellule du tableau des fréquences théoriques représente le nombre de cas attendus sous l'hypothèse nulle dans le tableau de contingence, basé sur les distributions marginales des genres et des niveaux de CDR.

Fréquences théoriques attendues :

```
Fréquences théoriques attendues :
[[119.68208092  19.84971098  62.46820809]
 [ 85.31791908  14.15028902  44.53179191]]
```

Les tableaux des fréquences théoriques sont les mêmes dans les deux cas car les deux méthodes utilisées pour les calculer,Cela démontre la cohérence des calculs et la validité des résultats obtenus.

Le test du chi-deux :

```
Test du chi-deux :
Statistique du test du chi-deux : 26.88213839490587
p-value : 1.4541788259515 25e-06
Degré de liberté : 2
```

```
Statistique du chi carré : 26.88213839490587
Valeur de p : 1.454178825951525e-06
On rejette H0 et on accepte H1, les deux vaiables sont dépendantes
```

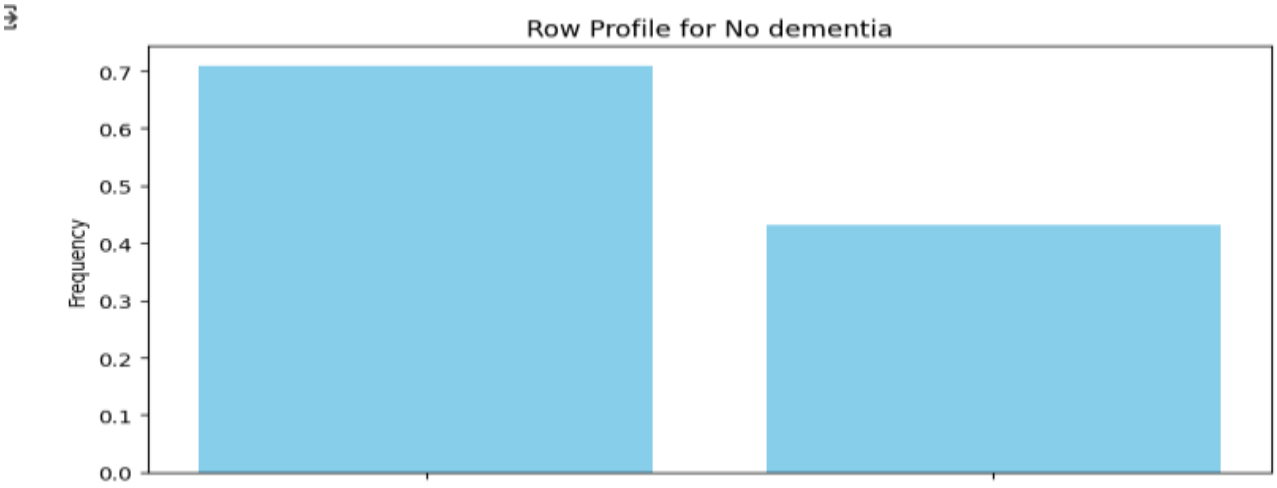
La p-value très faible (quasiment nulle) et la statistique du chi carré supérieure au seuil critique indiquent un rejet de l'hypothèse nulle d'indépendance entre le genre et le niveau de démence. Cela suggère fortement une association significative entre ces variables dans le tableau de contingence

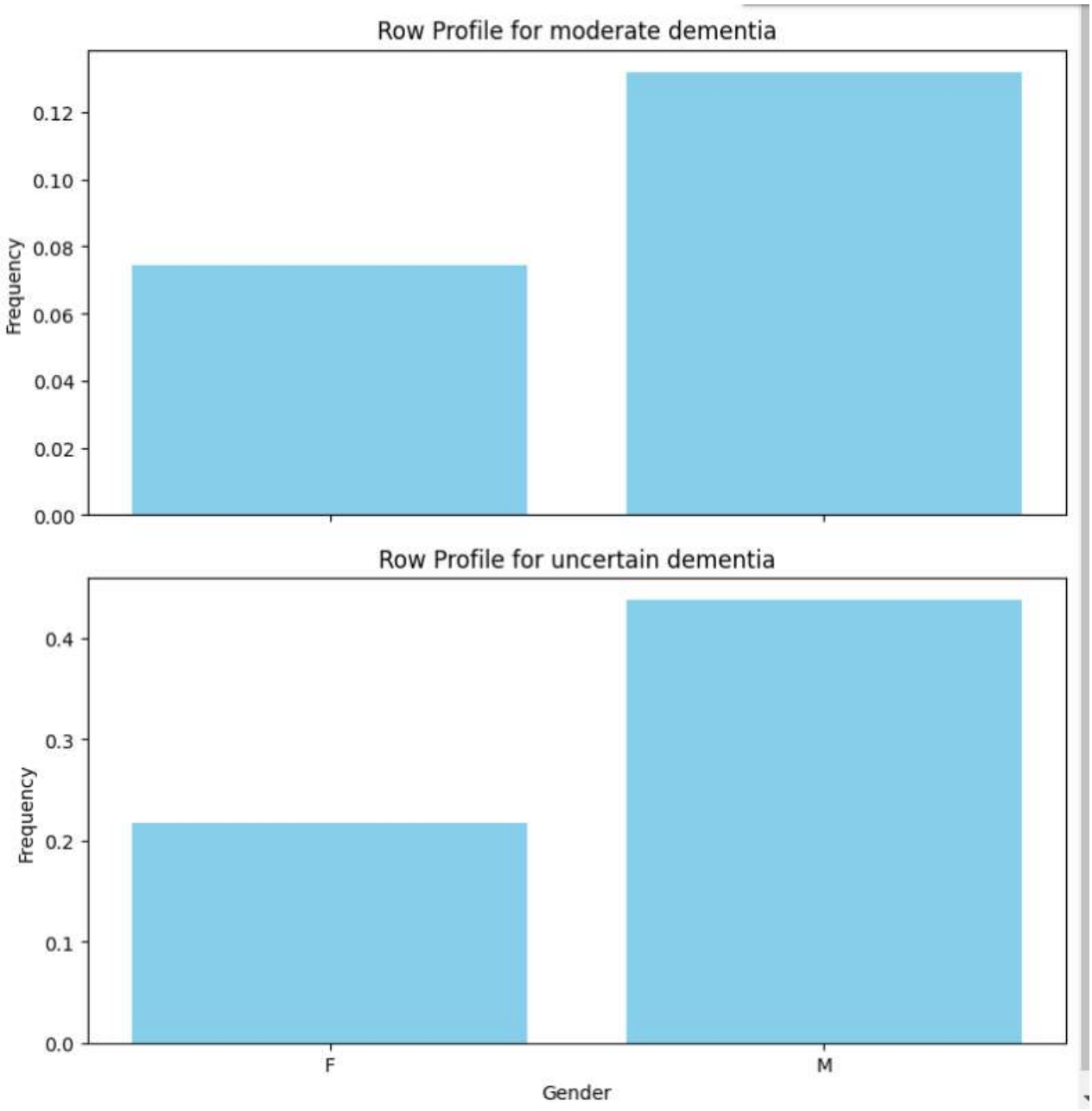
Distributions marginales :

```
Vecteur de poids des lignes (distribution marginale des lignes) :
 F    1.549952
M    1.450048
dtype: float64
Vecteur de poids des colonnes (distribution marginale des colonnes) :
 No dementia           1.0
moderate dementia     1.0
uncertain dementia    1.0
dtype: float64
```

Les vecteurs de poids des lignes et des colonnes représentent les distributions marginales des lignes et des colonnes dans la matrice de fréquences. Le vecteur de poids des lignes montre que la somme des fréquences pour les femmes (F) est égale à 1.55, tandis que pour les hommes (M), elle est de 1.45. De même, le vecteur de poids des colonnes montre que la somme des fréquences pour chaque catégorie de démence est égale à 1, ce qui est attendu dans une distribution marginale.

Plot des profiles lignes:





Pour la catégorie "No dementia" :

- Le profil-ligne montre que la fréquence de cette catégorie est plus élevée pour le genre féminin (F) que pour le genre masculin (M). Cela signifie qu'il y a une proportion plus élevée de cas sans démence parmi les femmes que parmi les hommes.

Pour la catégorie "moderate dementia" :

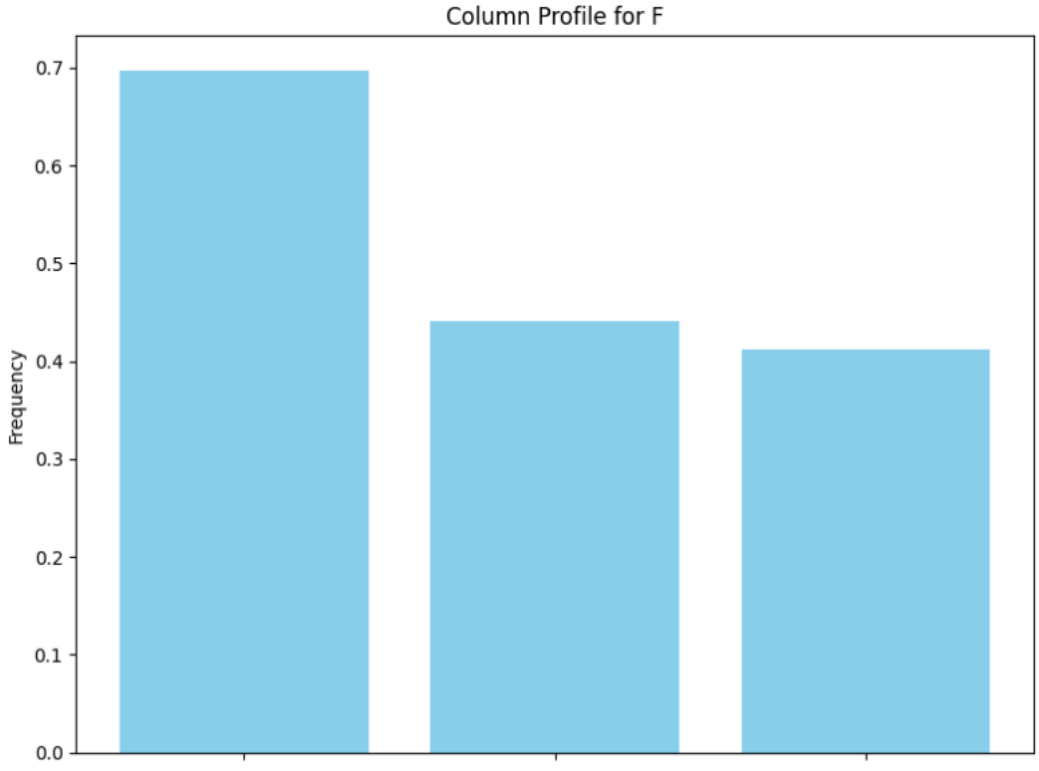
- Le profil-ligne indique également une fréquence plus élevée chez les femmes que chez les hommes, bien que la différence soit moins marquée que pour la catégorie "No dementia".

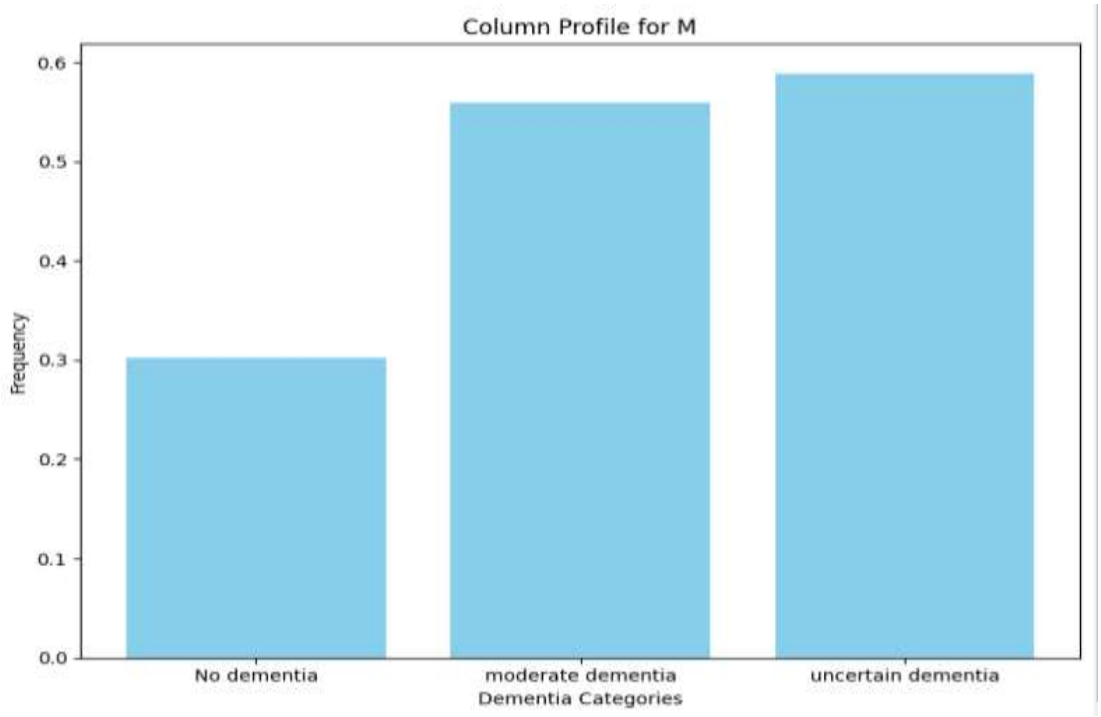
Pour la catégorie "uncertain dementia" :

- Contrairement aux deux premières catégories, le profil-ligne montre une fréquence légèrement plus élevée chez les hommes que chez les femmes, mais la différence n'est pas significative.

Globalement, les profils-lignes mettent en évidence des variations dans la fréquence des différents niveaux de démence entre les genres féminin et masculin. Cela pourrait indiquer une certaine association entre le genre et le niveau de démence .

Plot des profiles colonnes :





Pour le genre "F" (femme) :

- Le profil-colonne montre que la fréquence de chaque catégorie de démence est la plus élevée pour "No dementia", suivie par "moderate dementia" et enfin "uncertain dementia". Cela indique que parmi les femmes, il y a une proportion plus élevée de cas sans démence que de cas de démence modérée, et une proportion plus élevée de cas de démence modérée que de cas de démence incertaine.

Pour le genre "M" (homme) :

- Le profil-colonne montre un schéma similaire à celui des femmes, avec une fréquence décroissante de "No dementia" à "moderate dementia" à "uncertain dementia". Cependant, la proportion de chaque catégorie de démence diffère légèrement entre les genres, avec par exemple une fréquence plus élevée de démence modérée chez les hommes par rapport aux femmes.

Ces profils-colonnes mettent en lumière les différences de fréquence des différentes catégories de démence entre les genres féminin et masculin

Distribution conditionnelle en ligne :

```
Profil de ligne moyen :
 No dementia            1.0
moderate dementia       1.0
uncertain dementia      1.0
dtype: float64
Profil de colonne moyen :
 F    1.549952
M     1.450048
dtype: float64
```

Ces résultats indiquent que le profil de ligne moyen pour chaque catégorie de démence (No dementia, moderate dementia, uncertain dementia) est égal à 1.0. Cela signifie que chaque catégorie de démence représente 100% des observations dans le profil de ligne moyen, ce qui est cohérent car ces valeurs sont calculées à partir de la distribution conditionnelle en ligne, où chaque ligne représente une catégorie de démence.

Quant au profil de colonne moyen, il indique que la distribution marginale des colonnes pour les sexes (F et M) est légèrement déséquilibrée. En moyenne, il y a environ 1.55 observations par ligne pour le sexe féminin (F) et environ 1.45 observations par ligne pour le sexe masculin (M). Cela suggère un léger déséquilibre dans la répartition des sexes dans les données, mais il est important de noter que ces valeurs ne sont pas significativement différentes et que d'autres analyses pourraient être nécessaires pour comprendre pleinement les relations entre les variables.

La distance de chi2 entre deux modalités :

```
Distance de chi2 entre les deux modalités ( male et female) : 0.12851778594388588
```

La distance de chi2 entre les deux modalités (male et female) est calculée comme étant environ 0.129. Cette mesure de distance de chi2 est utilisée pour évaluer la différence entre les profils de lignes pour les deux modalités (male et female). Plus la distance de chi2 est grande, plus les profils de lignes pour les deux modalités sont différents. Dans ce cas, une distance de 0.129 suggère une différence relativement faible entre les profils de lignes pour male et female, ce qui indique une similitude relative dans la distribution des catégories de démence entre les deux groupes de sexe.

III. CONCLUSION :

En conclusion, ce projet a impliqué une analyse approfondie d'un ensemble de données. L'ensemble de données, contenant des caractéristiques liées à la maladie d'Alzheimer, a subi des étapes de prétraitement telles que la réduction de la multicolinéarité grâce à l'Analyse en Composantes Principales (ACP)

En utilisant l'algorithme K-means, nous avons réussi à regrouper les individus en clusters distincts, mettant en évidence des tendances et des différences significatives entre les groupes.

De plus, grâce à l'Analyse Factorielle des Correspondances (AFC), nous avons pu visualiser et interpréter efficacement les associations entre les variables qualitatives. En intégrant les résultats, bien que nous ayons identifié une association significative entre le genre et les profils de démence, les différences observées ne sont pas très marquées dans notre échantillon.

Ces analyses offrent des perspectives importantes pour une meilleure compréhension de la maladie d'Alzheimer et ouvrent la voie à de futures recherches et interventions cliniques.