

UNIVERSITÉ NATIONALE DU VIETNAM  
INSTITUT FRANCOPHONE  
INTERNATIONAL

---



FOUILLE DE DONNÉES  
RAPPORT DES TRAVAUX

Données utilisées : "**Infarctus**"

Juin 2019

Étudiants du Groupe 9 :

**Mike Arley MORIN**

**Afi Elolo Gisèle DEKPE**

Professeur : **Nguyen Thi Minh Huyen**

P23 RSC, Année Académique : 2019 - 2020

# Table des matières

<b>Introduction</b>	<b>6</b>
<b>1 Le choix du jeu de données</b>	<b>7</b>
<b>2 Description du jeu de données et résumé des données</b>	<b>7</b>
<b>3 Analyse exploratoire des variables et des paires de variables</b>	<b>8</b>
3.1 Étude d'une variable . . . . .	8
3.1.1 Cas d'une variable quantitative . . . . .	8
3.1.2 Cas d'une variable qualitative . . . . .	12
3.2 Étude des paires de variables . . . . .	16
3.2.1 Cas de deux (02) variables quantitatives . . . . .	16
3.2.2 Cas de deux (02) variables qualitatives . . . . .	18
3.2.3 Cas de deux (02) variables dont l'une quantitative et l'autre qualitative	20
<b>4 Analyse factorielle du jeu de données</b>	<b>21</b>
4.1 Analyse en composantes principales . . . . .	22
4.1.1 Les valeurs propres . . . . .	22
4.1.2 Corrélation entre les variables et les axes principaux . . . . .	24
4.1.3 Plans factoriels . . . . .	25
4.1.4 Cercle des corrélations . . . . .	25
4.2 Analyse factorielle des correspondances multiples - ACM . . . . .	26
4.2.1 Les valeurs propres . . . . .	27

4.2.2	Coordonnées des catégories des variables . . . . .	28
4.2.3	Contribution des variables aux dimensions . . . . .	29
4.2.4	Spécification dans l'ACM . . . . .	31
4.3	Analyse factorielle des données mixtes- AFDM . . . . .	31
4.3.1	Le tableau des valeurs propres . . . . .	32
4.3.2	Le tableau des coordonnées . . . . .	33
4.3.3	Le tableau des corrélations . . . . .	33
4.3.4	Le tableau des moyennes conditionnelles . . . . .	34
4.3.5	Représentation graphique des individus . . . . .	35
4.3.6	Le cercle de corrélation . . . . .	36
4.3.7	Représentation des moyennes conditionnelles . . . . .	37
<b>5</b>	<b>Classification - Clustering</b>	<b>37</b>
5.1	La classification Ascendante hiérarchique . . . . .	38
5.1.1	Le dendrogramme . . . . .	38
5.1.2	Le Graphique 3D combinant la classification hiérarchique et le plan des facteurs . . . . .	40
5.1.3	Variables quantitatives décrivant le plus chaque cluster . . . . .	41
5.1.4	Axes principaux associés aux clusters . . . . .	42
5.2	La méthode de K-MEANS . . . . .	43
5.2.1	Estimation du nombre optimal de clusters . . . . .	43
5.2.2	Classement des observations . . . . .	44
<b>Conclusion</b>		<b>46</b>

# Table des figures

1	Table des variables . . . . .	7
2	Tableau des indicateurs statistiques des variables quantitatives . . . . .	8
3	Histogramme de la variable Age . . . . .	9
4	Histogramme de la variable Poids . . . . .	9
5	Histogramme de la variable Taille . . . . .	10
6	Histogramme de la variable Imc . . . . .	10
7	Tableau des indicateurs statistiques des variables qualitatives . . . . .	12
8	Diagramme en secteur de la variable Infarct . . . . .	13
9	Diagramme en secteur de la variable CO . . . . .	13
10	Diagramme en secteur de la variable Tabac . . . . .	14
11	Diagramme en secteur de la variable ATCD . . . . .	14
12	Diagramme en secteur de la variable HTA . . . . .	15
13	Graphe des nuages . . . . .	17
14	Diagramme de dispersion de la corrélation entre le Poids et l'IMC . . . . .	18
15	Tableau de contingence "Infarct" et "Co" . . . . .	19
16	Résultat du test de Khi-deux sur les variables qualitatives . . . . .	19
17	Illustration de l'analyse des variances sur des variables de types diverses . . . . .	21
18	Table des valeurs propres . . . . .	23
19	Histogramme des valeurs propres . . . . .	23
20	Contribution des variables aux axes principaux . . . . .	24
21	Plan factorielle sur le premier axe . . . . .	25

22	Graphique de corrélation des variables : Cercle des corrélations . . . . .	26
23	Tableau des valeurs propres . . . . .	27
24	Histogramme des valeurs propres . . . . .	27
25	Corrélation des variables par rapport aux axes ( dimensions) . . . . .	28
26	Coordonnées des catégories des variables . . . . .	29
27	Contribution des variables aux dimensions . . . . .	29
28	Illustration - Contribution des variables aux dimensions . . . . .	30
29	Illustration dans l'ACM . . . . .	31
30	Tableau des valeurs propres . . . . .	32
31	Tableau des coordonnées . . . . .	33
32	Tableau des corrélations . . . . .	34
33	Tableau des moyennes conditionnelles . . . . .	34
34	La projection des individus . . . . .	35
35	Le cercle de corrélation . . . . .	36
36	Illustration des moyennes conditionnelles . . . . .	37
37	Le dendrogramme issu de notre classification . . . . .	38
38	Le plan facteur issu de notre classification . . . . .	39
39	Le Graphique 3D : classification hiérarchique & plan des facteurs . . . . .	40
40	Variables quantitatives décrivant le plus chaque cluster . . . . .	41
41	Axes principaux associés aux clusters . . . . .	42
42	Détermination du nombre k de clusters . . . . .	44
43	Visualisation graphique de la méthode K-Means . . . . .	45

# Introduction

La fouille de données est une approche analytique de données, adaptée et utilisée dans un large nombre de domaine d'activités. C'est une discipline qui vise à extraire les informations pertinentes d'un grand ensemble de données. Tout l'enjeu est de réussir à préparer, manipuler et analyser les données dans l'optique de les transformer en connaissance actionnable et en outil d'aide à la décision pour bon nombre de domaines de la vie active. Dans le but de nous imprégner dans le sujet des fouilles de données, il nous a été demandé de choisir un jeu de données sur lequel nous aurons à travailler en binôme. Pour ce faire, nous avons choisi le jeu de données "**Infarctus**". Le dataset choisi, traite des données issues d'une enquête dont le but était d'évaluer l'existence d'un risque d'infarctus du myocarde chez les femmes qui utilisent ou ont utilisé des contraceptifs oraux, associé à d'autres facteurs. L'étude a été menée auprès de 148 femmes ayant eu un infarctus du myocarde (cas) et 300 femmes n'en ayant pas eu (témoins). Le facteur d'exposition principal est la prise de contraceptifs oraux, les autres facteurs recueillis sont : l'âge, le poids, la taille, la consommation de tabac, l'hypertension artérielle, les antécédents familiaux de maladies cardio-vasculaires. Ce document fait office du rapport des premiers traitements à effectuer sur les données. Dans les lignes suivantes, nous allons faire la description des données puis une analyse exploratoire des variables et des paires de variables.

# 1 Le choix du jeu de données

Un jeu de données est un ensemble de valeurs (ou données) où chaque valeur est associée à une variable (ou attribut) et à une observation. Une variable décrit l'ensemble des valeurs décrivant le même attribut et une observation contient l'ensemble des valeurs décrivant les attributs d'une unité (ou individu statistique). Pour ce travail, nous devrons choisir un jeu de données qu'on utilisera également pour les autres travaux pratiques de la Fouille de données à venir. Dans notre cas, nous avons choisi « **"Infarctus"** », comme cité plus haut parce qu'il contient non seulement des valeurs d'entrées mais aussi il contient une valeur de sortie. Notre travail va porter sur les données du fichier infarctus.csv.

## 2 Description du jeu de données et résumé des données

L'ensemble de données « Infarctus » comporte 9 variables dont 5 variables qualitatives et 4 variables quantitatives ; avec pour variable de sortie "**Infarct**" :

N°	Type	Variables	Description	Unité ou Codage
1	Qualitative	<b>Infarct</b>	Infarctus du Myocarde	0: témoins 1:Cas
2		<b>Co</b>	Prise de contraceptifs oraux	0: Jamais 1:Oui
3		<b>Tabac</b>	Consommation de tabac	0: Non 1: Fumeuse 2: Ancienne fumeuse
4		<b>Atcd</b>	Antécédents familiaux de maladie cardio-vasculaire	0: Nom 1: Oui
5		<b>Hta</b>	Hypertension artérielle	0: Non 1: Oui
6	Quantitative	<b>Age</b>	Age	Années
7		<b>Poids</b>	Poids	Kg
8		<b>Taille</b>	Taille	Cm
9		<b>Imc</b>	Indice de masse corporelle	Kg/m <sup>2</sup>

FIGURE 1 – Table des variables

Source : <http://www.biostatisticien.eu/springeR/jeuxDonnees5.html>

Les variables étant ainsi présentées, nous allons passer à l'étude des variables avec le logiciel de Fouilles de Données **R**. Rappelons que **R** est un logiciel gratuit et open-source qui fonctionne sous Linux et bien d'autres plateformes, particulièrement performant pour la manipulation de données, le calcul et l'affichage de graphiques.

### 3 Analyse exploratoire des variables et des paires de variables

Dans cette section, nous allons étudier les variables de façon individuelle d'abord puis pris en paire.

#### 3.1 Étude d'une variable

##### 3.1.1 Cas d'une variable quantitative

Nous présentons dans le tableau des indicateurs statistiques(Tendance centrale et Dispersion) ci-dessous l'étude des variables quantitatives de notre jeu de données et à travers les graphique suivants :

(i) *Le tableau des indicateurs statistiques des variables quantitatives*

Variables		Age	Poids	Taille	Imc
Tendance centrale	<b>Effectif</b>	448	448	448	448
	<b>Moyenne</b>	45.61	66.44	165.14	24.20
	<b>Médiane</b>	44	64	166	23.03
	<b>Mode</b>	42	45	170	16.35
Dispersion	<b>Min</b>	15	33	138	11.36
	<b>Max</b>	100	128	184	47.78
	<b>Étendue</b>	15 - 100	33 - 128	138 - 184	11.36 - 47.78
	<b>Variance</b>	261.9115	326.31	65.79	51.04
	<b>Écart-type</b>	16.18368	18.06	8.11	7.14

FIGURE 2 – Tableau des indicateurs statistiques des variables quantitatives

(ii) *Les histogrammes des variables quantitatives*

— Histogramme : Age

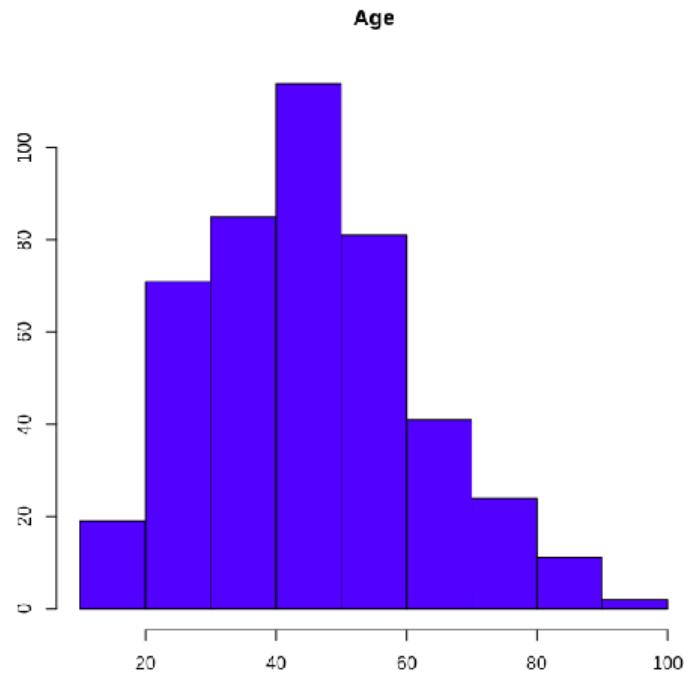


FIGURE 3 – Histogramme de la variable Age

— Histogramme : *Poids*

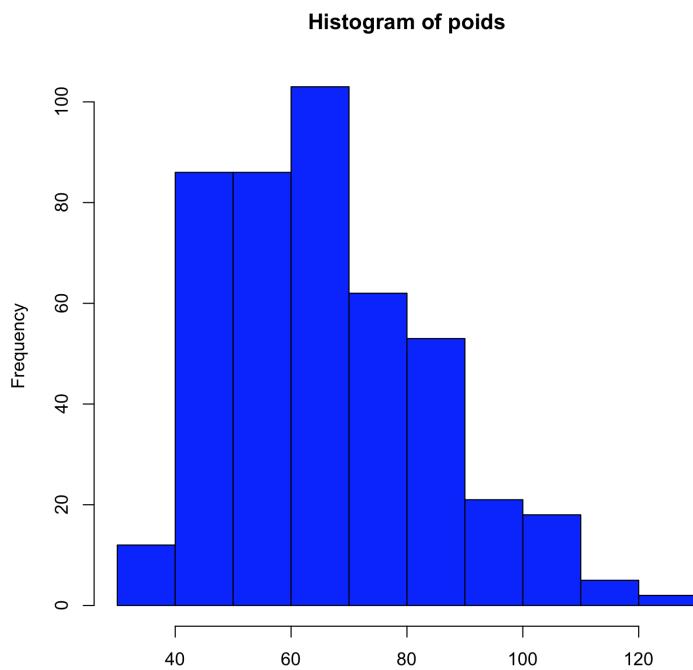


FIGURE 4 – Histogramme de la variable Poids

— Histogramme : *Taille*

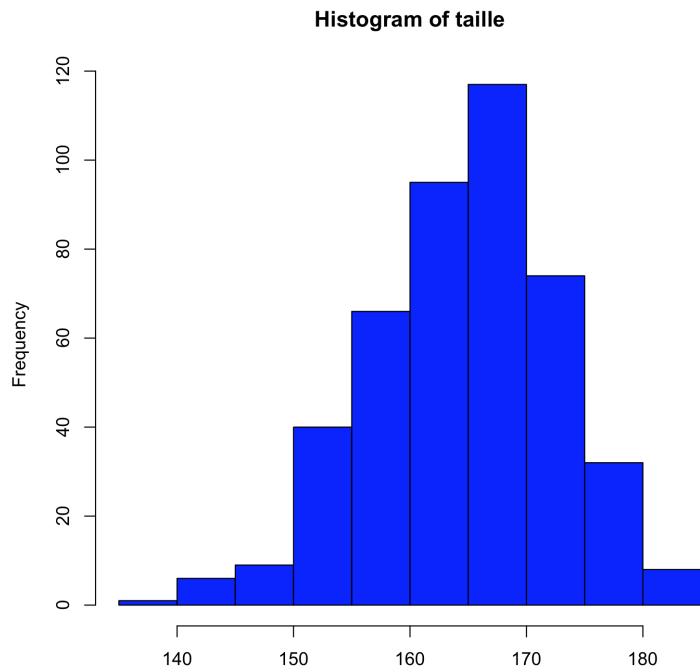


FIGURE 5 – Histogramme de la variable Taille

— Histogramme : *Imc*

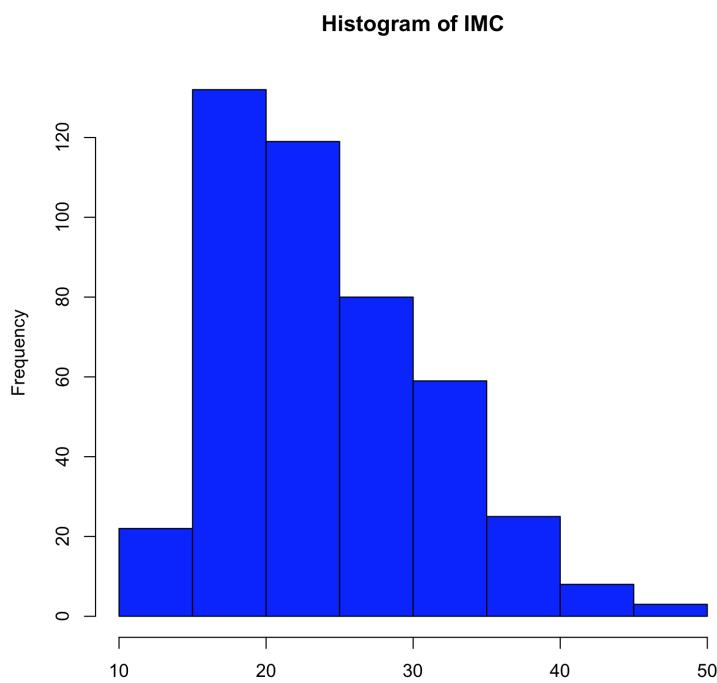


FIGURE 6 – Histogramme de la variable Imc

(iii) Interprétation :

- Pour l'*Age* : les traitements nous montre les informations suivantes :
  - Indicateurs statistiques de Tendance centrale :

- la moyenne = 45,61 : ainsi la moyenne d'âge de la population étudiée est estimée à 45 ans ; - la médiane = 44 : la moitié de la population étudiée a plus de 44 ans tandis que l'autre moitié a moins de 44 ans ; - le mode est 42 : la majorité a 42 ans ;
- Indicateurs statistiques de Dispersion : - l'écartype = 16,18 ; - l'étendue = 85 : la différence d'âge maximale est de 85 ans ; - La dispersion interquartile (variance) est de 261.91.

- Pour le *Poids* : les traitements nous montre les informations suivantes :
  - Indicateurs statistiques de Tendance centrale :
  - la moyenne = 66.44 KG ; - la médiane = 64 : la moitié de la population étudiée pèse plus de 44 ans tandis que l'autre moitié a moins de 44 ans ; - le mode est 45 : la majorité a 45KG comme poids ;
  - Indicateurs statistiques de Dispersion : - l'écartype = 18.06 ; - l'étendue = 95 : la différence de poids maximale est de 95 Kg ; - La dispersion interquartile (variance) est de 326.31.
- Pour la *Taille* : les traitements nous montre les informations suivantes :
  - Indicateurs statistiques de Tendance centrale :
  - la moyenne = 165.14 : ainsi la taille moyenne de la population étudiée est estimée à 1m 65 cm ; - la médiane = 166 : la moitié de la population étudiée mesure plus de 1m 66 tandis que l'autre moitié moins de 1m 66 ; - le mode est 170 : la majorité 1m 70 ;
  - Indicateurs statistiques de Dispersion : - l'écartype = 8.11 ; - l'étendue = 46 : la différence de taille maximale est de 0.46 m ; - La dispersion interquartile (variance) est de 65.79.
- Pour l'*IMC* : les traitements nous montre les informations suivantes :
  - Indicateurs statistiques de Tendance centrale :
  - la moyenne = 24.20 : ainsi la moyenne d'indice de masse corporelle de la population étudiée est estimée à 24.2 Kg/m ; - la médiane = 23.03 : la moitié de la population étudiée a plus de 23.03 Kg/m tandis que l'autre moitié a moins de 23.03 Kg/m ; - le mode est 16.35 Kg/m : la majorité a 16.35 Kg/m ;
  - Indicateurs statistiques de Dispersion : - l'écartype = 7.14 Kg/m ; - l'étendue =

36.42 Kg/m : la différence d'indice de masse corporelle maximale est de 36.42 Kg/m ;

- La dispersion inter-quartile (variance) est de 51.04 Kg/m.

### 3.1.2 Cas d'une variable qualitative

L'étude des variables qualitatives est également résumée dans le tableau ci-dessous et à travers les graphiques suivants :

(i) *Le tableau des indicateurs statistiques des variables qualitatives*

Variables	Infarct		Co		Tabac			ATCD		HTA	
Valeurs	0	1	0	1	0	1	2	0	1	0	1
Fréquence	300	148	249	199	215	135	98	395	53	290	158
Fréquence cumulée	<b>448</b>										
Pourcentage	66.96	33.04	55.58	44.41	47.99	30.13	21.87	88.17	11.83	64.73	35.26
Pourcentage cumulée	<b>100</b>										

FIGURE 7 – Tableau des indicateurs statistiques des variables qualitatives

(ii) *Les diagrammes en secteur des variables qualitatives*

— Diagramme en secteur : *Infarct*

**INFARCT**

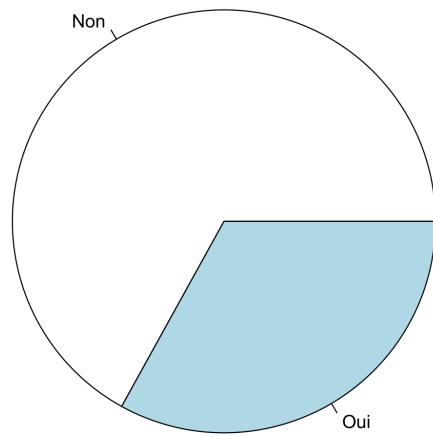


FIGURE 8 – Diagramme en secteur de la variable Infarct

— Diagramme en secteur : *Co*

**CO**

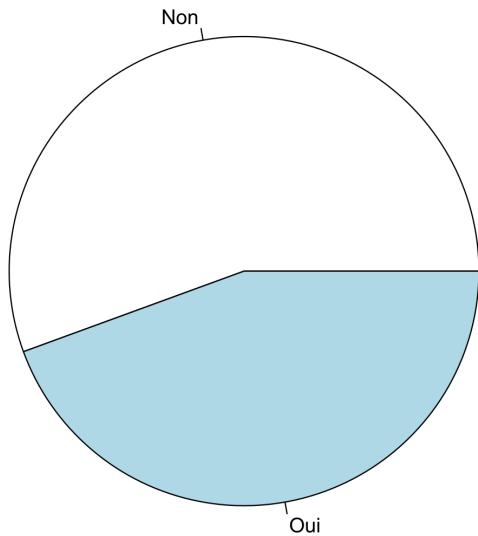


FIGURE 9 – Diagramme en secteur de la variable CO

— Diagramme en secteur : *Tabac*

**TABAC**

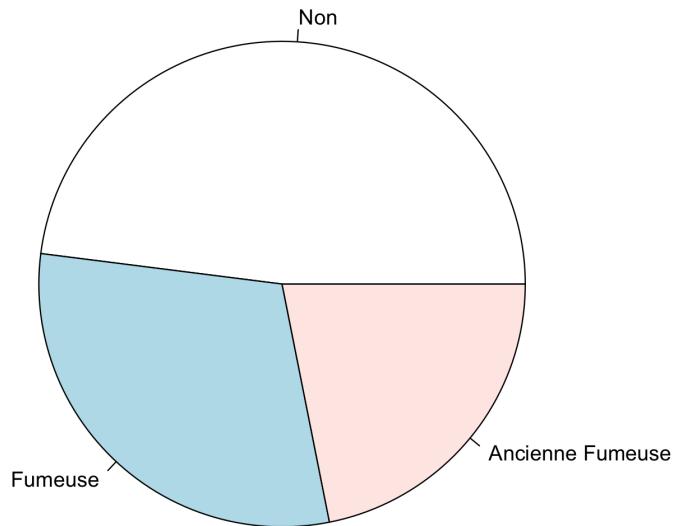


FIGURE 10 – Diagramme en secteur de la variable Tabac

— Diagramme en secteur : *ATCD*

**ATCD**

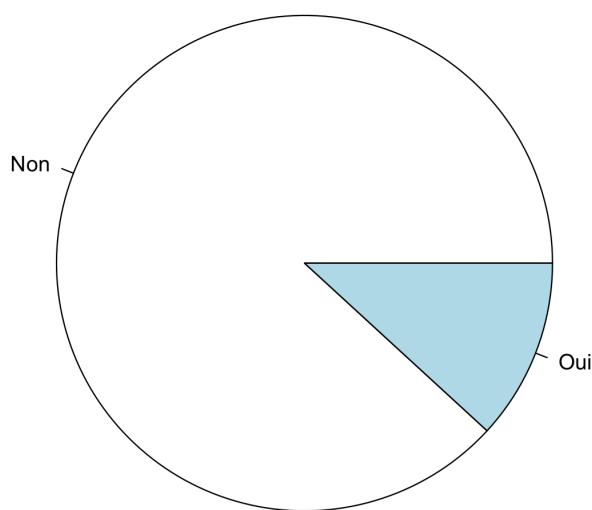


FIGURE 11 – Diagramme en secteur de la variable ATCD

— Diagramme en secteur : *HTA*

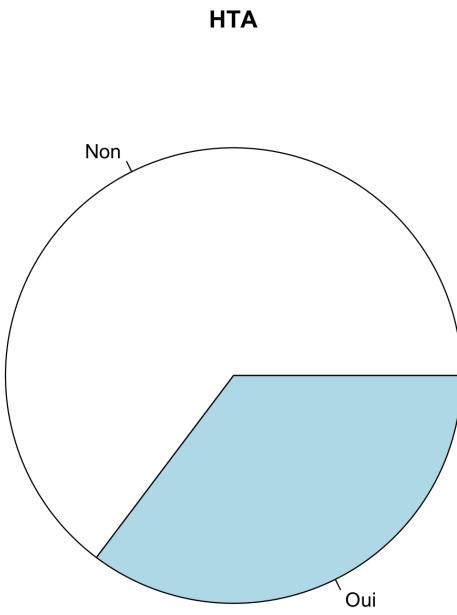


FIGURE 12 – Diagramme en secteur de la variable HTA

(iii) Interprétation :

- Pour la variable *Infarct* : d'après le résultat des calculs et de la représentation en secteur ci-dessus, le mode de la variable « Infarct » est « 0 » puisque parmi les valeurs correspondantes à l'attribut Infarct, c'est « 0 » qui a la plus forte proportion d'invidus (**66.96%**). La plupart de la population étudiée est donc **témoins d'infarctus du myocarde**.
- Pour la variable *Co* : d'après le résultat des calculs et de la représentation en secteur ci-dessus, le mode de la variable « Co » est « 0 » puisque parmi les valeurs correspondantes à l'attribut Co, c'est « 0 » qui a la plus forte proportion d'invidus (**55.58%**). La plupart de la population étudiée **ne prend donc pas de contraceptifs oraux**.
- Pour la variable *Tabac* : d'après le résultat des calculs et de la représentation en secteur ci-dessus, le mode de la variable « Tabac » est « 0 » puisque parmi les valeurs correspondantes à l'attribut Tabac, c'est « 0 » qui a la plus forte proportion d'invidus (**47.99%**). La plupart de la population étudiée **ne fume pas**.
- Pour la variable *ATCD* : d'après le résultat des calculs et de la représentation en secteur ci-dessus, le mode de la variable « ATCD » est « 0 » puisque parmi les valeurs correspondantes à l'attribut ATCD, c'est « 0 » qui a la plus forte proportion d'invidus (**88.17%**). La plupart de la population étudiée **n'a pas d'antécédent**

familial de maladie cardio-vasculaire.

- Pour la variable *HTA* : d'après le résultat des calculs et de la représentation en secteur ci-dessus, le mode de la variable « HTA » est « 0 » puisque parmi les valeurs correspondantes à l'attribut HTA, c'est « 0 » qui a la plus forte proportion d'invidus (**64.73%**). La plupart de la population étudiée **ne souffre pas d'hypertension artérielle**.

## 3.2 Étude des paires de variables

Ici, il s'agit de la relation entre deux paires de variables, selon les types puis entre les deux différents types.

### 3.2.1 Cas de deux (02) variables quantitatives

Pour ce faire, nous avons calculé la corrélation entre chaque paire de variables. Le résultat donne ce qui suit :

```
> cor(quantitative)
      AGE       POIDS      TAILLE       IMC
AGE  1.00000000  0.09402599 -0.07193286  0.1127975
POIDS  0.09402599  1.00000000 -0.01427630  0.8866215
TAILLE -0.07193286 -0.01427630  1.00000000 -0.3317689
IMC   0.11279752  0.88662152 -0.33176888  1.0000000
```

Nous remarquons que la valeur trouvée pour la corrélation entre le *Poids* et l'*IMC* est la plus élevée. Le graphe de nuage de point correspondant est le suivant :

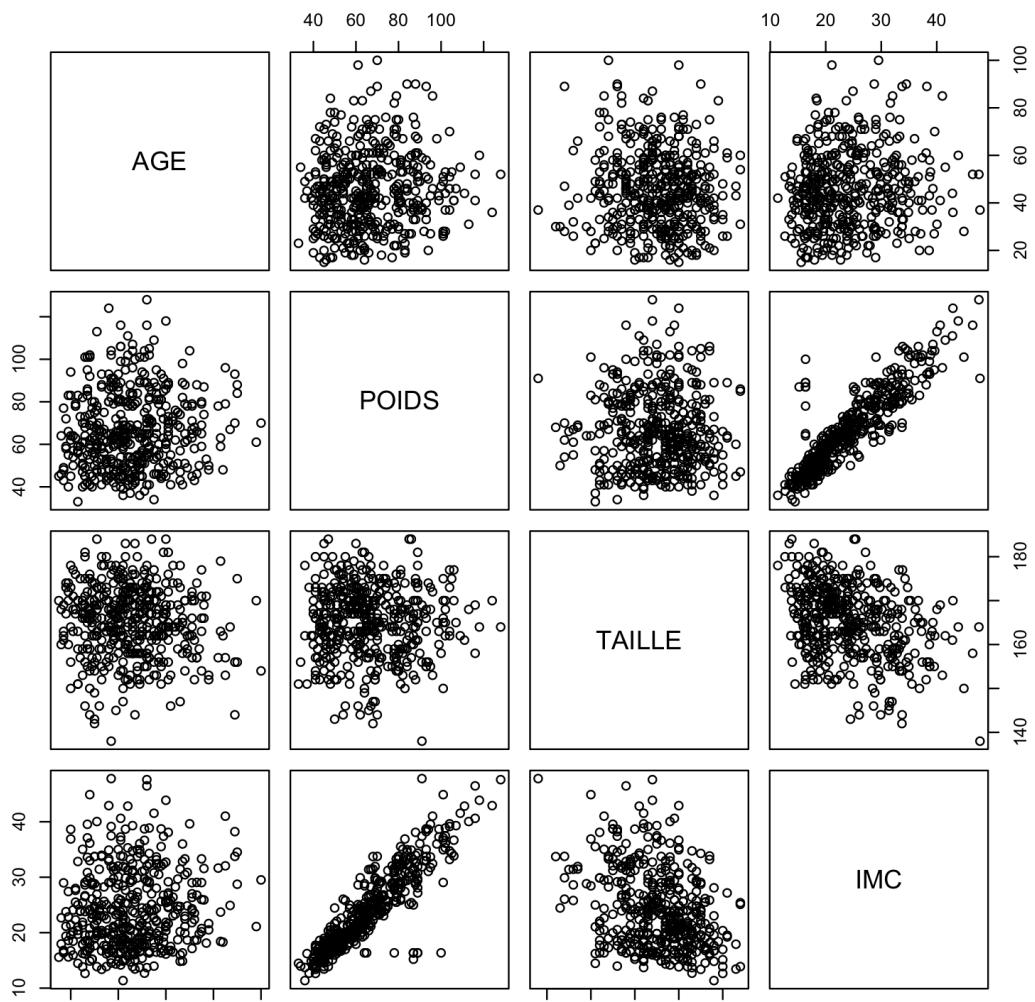


FIGURE 13 – Graphe des nuages

Nous obtenons ce diagramme de dispersion montrant la corrélation entre les deux variables (Poids et IMC) :

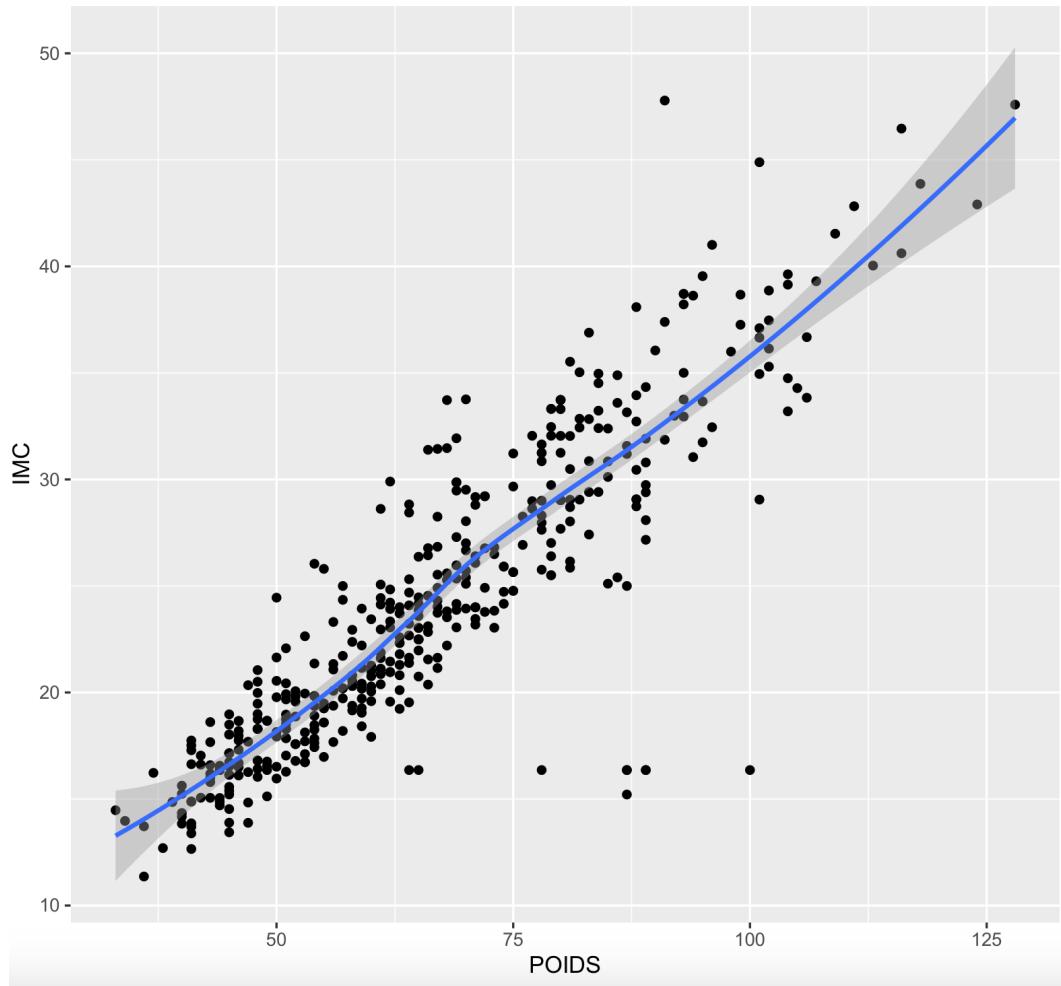


FIGURE 14 – Diagramme de dispersion de la corrélation entre le Poids et l’IMC

**Interprétation :** Le constat est que le nuage représenté est concentré autour de la droite de régression. Ce qui prouve qu'il y a une **forte corrélation entre les variables *poids* et *IMC***. Lorsque les valeurs de *poids* augmentent, celles d'*IMC* également augmentent pour la plupart.

### 3.2.2 Cas de deux (02) variables qualitatives

Nous allons chercher à trouver l'indépendance entre deux variables qualitatives. Nous allons effectuer un test khi-deux ( 2 ) à cet effet.

Les variables choisies sont l'infarctus du Myocarde "*Infarct*" et la prise de contraceptifs oraux "*Co*". L'objectif est de vérifier si les femmes qui prennent des contraceptifs sont sujettes à l'infarctus du myocarde.

**Hypothèse nulle (H0)** : La prise des contraceptifs oraux n'ont pas d'effet sur la contraction de l'infarctus du myocarde.

Si l'hypothèse est rejetée, alors il y a un lien entre la prise des contraceptifs et l'infarctus du myocarde. Avec R, nous avons fait un tableau de contingence avec les variables choisies ; puis déterminé la valeur **p-value** :  $2,2\text{exp}(-16)$  qui équivaut à **0.00000332075**

```
> table(data$CO,data$INFARCT, deparse.level = 2 )
  data$INFARCT
data$CO   0   1
  0 212 37
  1 88 111
~
```

FIGURE 15 – Tableau de contingence "Infarct" et "Co"

Le tableau de contingence montre que les femmes qui ne prennent pas des contraceptifs oraux sont celles qui ne sont pas touchées par l'infarctus du myocarde.

	khi <sup>2</sup>	dF	p-Value
<b>Infarct &amp; Co</b>	<b>81.876</b>	<b>1</b>	<b>&lt; 2.2e-16</b>
<b>Infarct &amp; Tabac</b>	<b>58.338</b>	<b>2</b>	<b>2.148e-13</b>
<b>Infarct &amp; ATCD</b>	<b>1.5408</b>	<b>1</b>	<b>0.2145</b>
<b>Infarct &amp; HTA</b>	<b>4.6921</b>	<b>1</b>	<b>0.0303</b>
<b>Co &amp; Tabac</b>	<b>20.808</b>	<b>2</b>	<b>3.032e-05</b>
<b>Co &amp; ATCD</b>	<b>3.8584e-30</b>	<b>1</b>	<b>1</b>
<b>Co &amp; HTA</b>	<b>2.3377</b>	<b>1</b>	<b>0.1263</b>
<b>Tabac &amp; ATCD</b>	<b>0.11813</b>	<b>2</b>	<b>0.9426</b>
<b>Tabac &amp; HTA</b>	<b>11.563</b>	<b>2</b>	<b>0.00308</b>
<b>ATCD &amp; HTA</b>	<b>5.7146</b>	<b>2</b>	<b>0.01682</b>

FIGURE 16 – Résultat du test de Khi-deux sur les variables qualitatives

**Interprétation du résultat** : khi-deux = 81,876 et p-value = 0.00000332075. p-value est très petite et inférieure au seuil significatif (0,05). Les variables présentent donc une association statistiquement significative. L'hypothèse H0 est rejetée.

**Conclusion :** Il y a donc un lien fort entre la prise des contraceptifs oraux et l'infarctus.

### 3.2.3 Cas de deux (02) variables dont l'une quantitative et l'autre qualitative

A ce niveau, nous allons faire une analyse de variance entre les variables quantitatives et la variable qualitative "Infarct" afin de déterminer là où il y a plus de lien. Nous posons une hypothèse pour chacune des variables quantitatives.

**Hypothèse nulle 1 (H0-1)** : L'âge n'a pas d'influence sur l'infarctus du myocarde.

**Hypothèse nulle 2 (H0-2)** : Le poids a une influence sur l'infarctus du myocarde.

**Hypothèse nulle 3 (H0-3)** : La taille a une influence sur l'infarctus du myocarde.

**Hypothèse nulle 4 (H0-4)** : L'IMC a une influence sur l'infarctus du myocarde.

Les résultats du test ANOVA se présentent comme suit :

Test d'ANOVA sur l'Age :

#### Analysis of Variance Table

```
Response: inf
          Df Sum Sq Mean Sq F value    Pr(>F)
factor(age)   2 2236547 1118273  94.672 < 2.2e-16 ***
Residuals   445 5256365    11812
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

\* Test d'ANOVA sur le Poids :

#### Analysis of Variance Table

```
Response: inf
          Df Sum Sq Mean Sq F value    Pr(>F)
factor(poids)  72 1196791    16622     0.99 0.5055
Residuals     375 6296121    16790
```

\* Test d'ANOVA sur la Taille :

### Analysis of Variance Table

Response: `inf`

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<code>factor(tay)</code>	78	1177613	15098	0.8821	0.7462
<code>Residuals</code>	369	6315299	17115		

\* Test d'ANOVA sur l'IMC :

### Analysis of Variance Table

Response: `inf`

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<code>factor(imc)</code>	41	583918	14242	0.8369	0.7534
<code>Residuals</code>	406	6908994	17017		

Récapitulatif des résultats

	Age	Poids	Taille	Imc
<code>Infarct</code>	2.2e-16	0.5055	0.7462	0.7534

FIGURE 17 – Illustration de l'analyse des variances sur des variables de types diverses

**Interprétation :** De toutes les valeurs calculées, la valeur de p liée à l'âge (**2.2e-16**) est la seule inférieure au seuil de signification(**0.05**). Nous pouvons rejeter l'hypothèse nulle 1 et garder les trois (03) autres.

## 4 Analyse factorielle du jeu de données

Pour réaliser cette partie de l'analyse, nous avons converti les variables qualitatives en variables numériques ; que nous utiliserons plupart pour la classification ou pour la prédiction. Dans notre jeu de données, les variables concernées sont *CO*, *TABAC*, *ATCD*, *HTA*.

Avec R, les étapes qui nous ont permis de réaliser ceci sont :

- la création d'un tableau qui contient les variables de valeur numérique : `tableauNumérique <- %> % select (5 :8), (5 :8)` pour la sélection de nos variables numériques

- la conversion des variables non numériques en numériques :
  - **nomDuDataset[ ,positionVariable] -> vecteurVariable**
  - **variableTemp1 <- factor(vecteurVariable)**
  - **variableTemp2 <- as.numeric(variableTemp1)**
- l'ajout de la variable vecteur dans le tableau des variables de valeur numérique : **tableauNumerique\$vecteurVariable <- vecteurVariable.numeric.**

## 4.1 Analyse en composantes principales

L'analyse en composantes principales (ACP) , ou principal component analysis (PCA) en anglais, permet d'analyser et de visualiser un jeu de données contenant des individus décrits par plusieurs variables quantitatives. Elle permet d'explorer des données avec plusieurs variables. Chaque variable pourrait être considérée comme une dimension différente. L'ACP réduit les dimensions d'une donnée multivariée à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'informations.

Pour notre cas de traitement sur le dataset **infarctus**, les lignes suivantes décrivent les traitements ACP effectués. L'objectif de cette analyse en composante principale est de partir d'un ensemble de données contenant 448 observations et 9 variables dont une de sortie, pour chercher à résumer l'information disponible à l'aide de variables synthétiques appelées composantes principales.

Le tableau des corrélation qui sera utilisée pour l'étude factorielle :

	CO	TABAC	AGE	POIDS	TAILLE	IMC	ATCD	HTA
CO	1.000000000	0.214831667	-0.24557914	-0.009033425	-0.02592647	0.002250073	0.006364894	-0.07693761
TABAC	0.214831667	1.000000000	-0.30562133	0.122223183	0.03105211	0.130097816	-0.001379484	-0.14554097
AGE	-0.245579138	-0.305621327	1.000000000	0.094025986	-0.07193286	0.112797516	-0.022409315	0.34920382
POIDS	-0.009033425	0.122223183	0.09402599	1.000000000	-0.01427630	0.886621517	0.150232545	0.18436998
TAILLE	-0.025926466	0.031052111	-0.07193286	-0.014276296	1.000000000	-0.331768877	0.041006714	0.02731272
IMC	0.002250073	0.130097816	0.11279752	0.886621517	-0.33176888	1.000000000	0.108626467	0.19211593
ATCD	0.006364894	-0.001379484	-0.02240932	0.150232545	0.04100671	0.108626467	1.000000000	0.12017451
HTA	-0.076937613	-0.145540971	0.34920382	0.184369979	0.02731272	0.192115930	0.120174506	1.000000000

### 4.1.1 Les valeurs propres

Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération. Ci-dessous les valeurs propres de la matrice des corrélations :

```

> eig.val
      eigenvalue variance.percent cumulative.variance.percent
Dim.1 1.97767504          49.441876                  49.44188
Dim.2 1.01596073          25.399018                  74.84089
Dim.3 0.94841350          23.710338                  98.55123
Dim.4 0.05795072          1.448768                  100.00000
>

```

FIGURE 18 – Table des valeurs propres

Les deux premières composantes principales représentent plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. Une autre méthode pour déterminer le nombre de composantes principales est de regarder le graphique des valeurs propres appelé **scree plot**, ci-dessous.

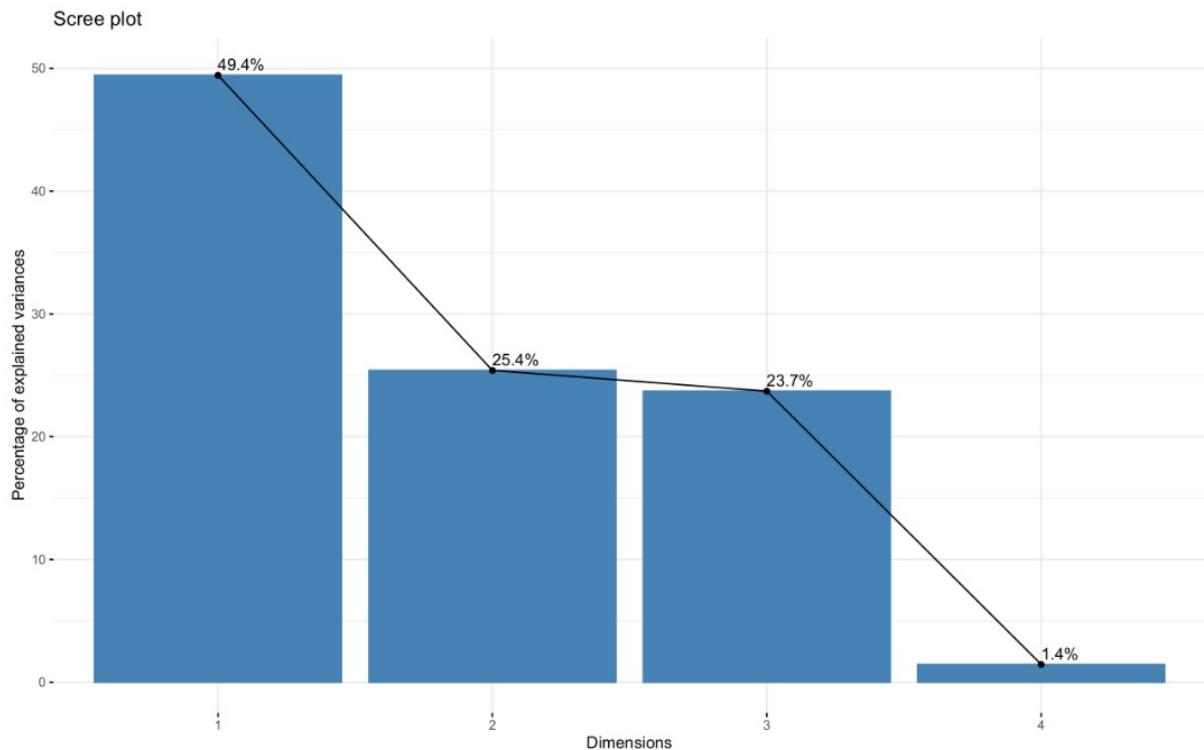


FIGURE 19 – Histogramme des valeurs propres

D'après le tableau précédent et l'histogramme présenté, l'inertie expliquée par le premier axe principal est **1.98**. En d'autres termes, la première composante principale explique à elle seule 49.44% (soit presque la moitié) de la variance totale. Dans notre analyse, les trois premières composantes principales expliquent **98.5%** de la variation. Du graphique ci-dessus, nous pourrions vouloir nous arrêter à la troisième composante principale. **98.5%** des informations (variances) contenues dans les données sont conservées par les trois premières composantes

principales. C'est un pourcentage acceptable ; nous continuons donc notre études avec ces trois (03) composantes.

#### 4.1.2 Corrélation entre les variables et les axes principaux

La seconde partie des résultats (voir figure 10) indique la corrélation des variables avec les axes factoriels.

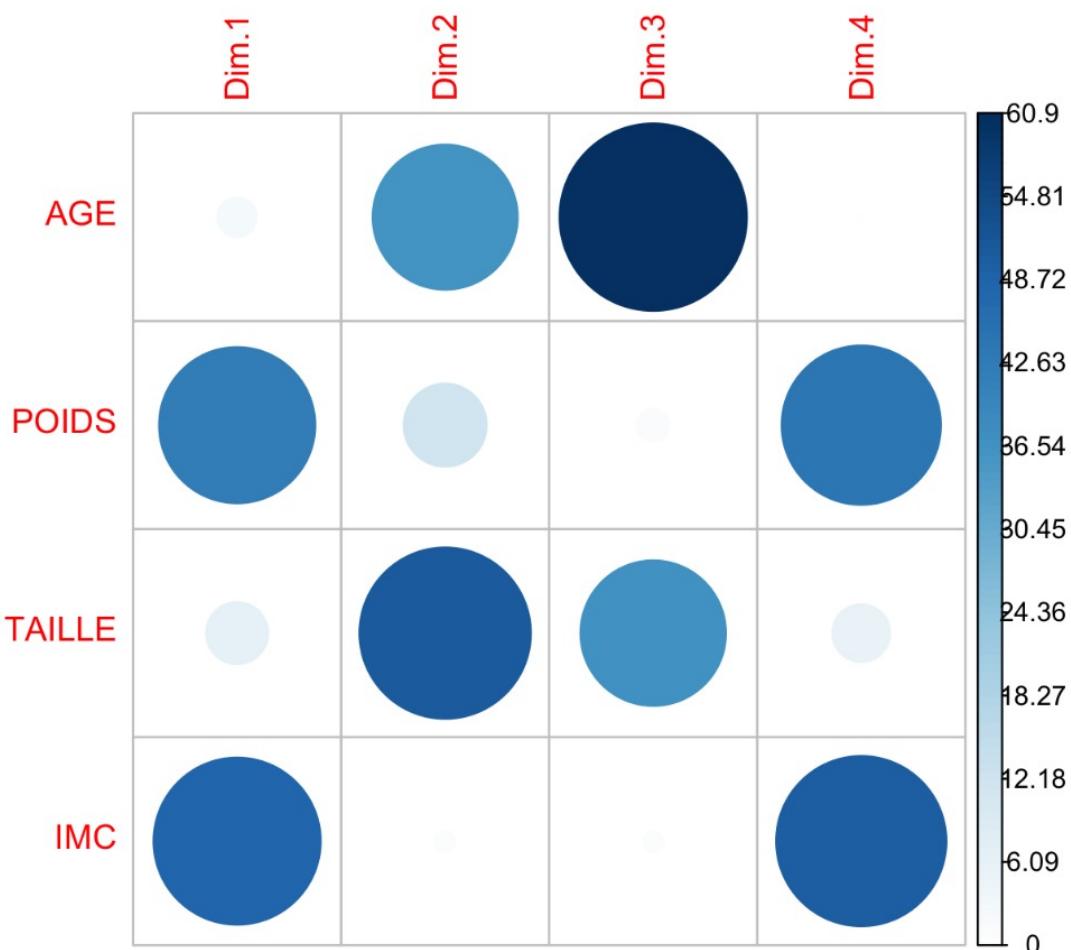


FIGURE 20 – Contribution des variables aux axes principaux

Plus la valeur de la contribution est importante, plus la variable contribue à la composante principale en question. Les variables corrélées sont les plus importantes pour expliquer la variabilité dans le jeu de données. Les variables qui ne sont pas en corrélation avec un axe ou qui sont corrélées avec les derniers axes sont des variables à faible apport et peuvent être supprimées pour simplifier l'analyse globale. De ce fait, nous pouvons dire que l'âge contribue

fortement à la dimension 3. L'IMC et la taille participe de façon remarquable (mais pas aussi fort que l'âge) respectivement aux dimensions 1 et 4 d'une part puis 2 d'autre part.

#### 4.1.3 Plans factoriels

Nous projetons les observations dans le premier plan factoriel afin de voir leur répartition.

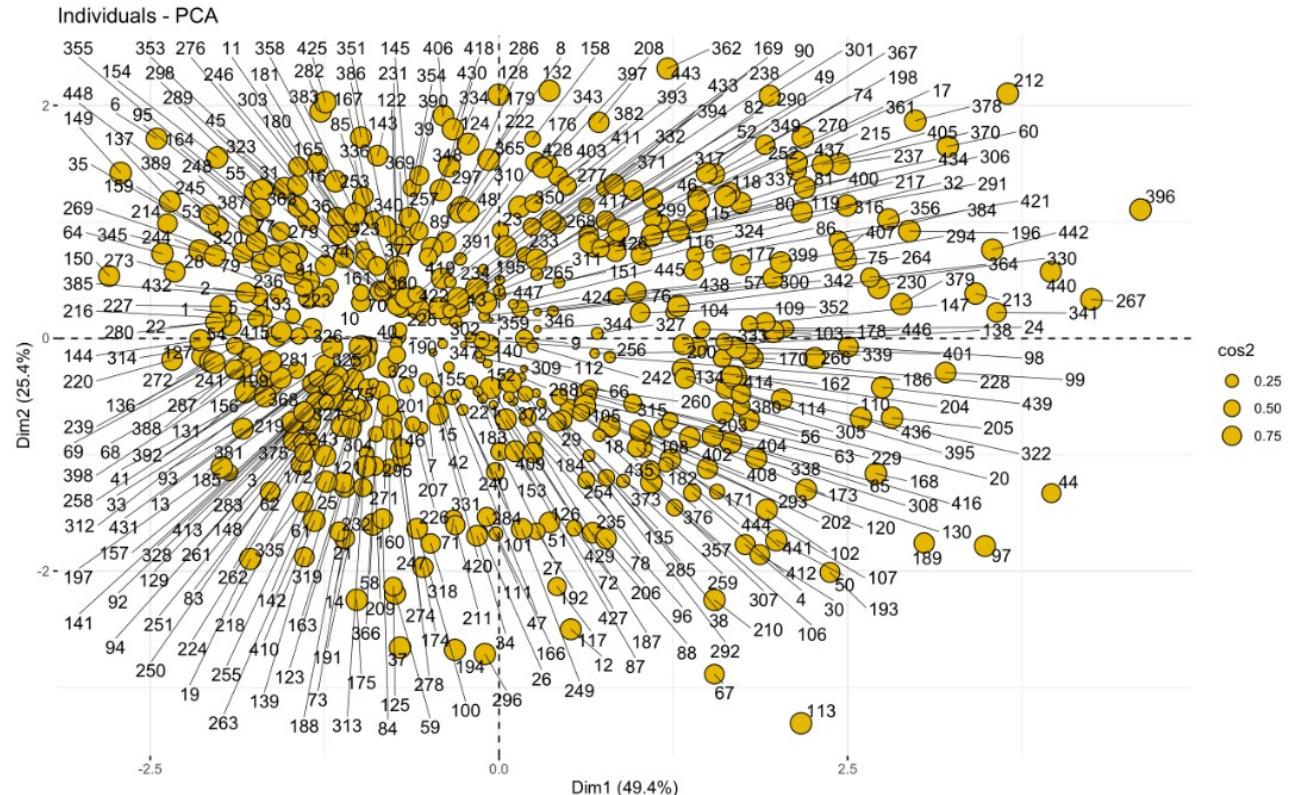


FIGURE 21 – Plan factoriel sur le premier axe

Les individus sont représentés par leur numéro dans le jeu de données (numéros/nombre affichés). Ceux qui sont similaires sont regroupés sur le graphique au travers d'une même forme. Une liste de matrices contenant tous les résultats pour les individus (coordonnées, corrélation entre individus et axes, cosinus-carré et contributions) est présentée en graphique.

#### 4.1.4 Cercle des corrélations

La corrélation entre une variable et une composante principale (PC) est utilisée comme coordonnées de la variable sur la composante principale.

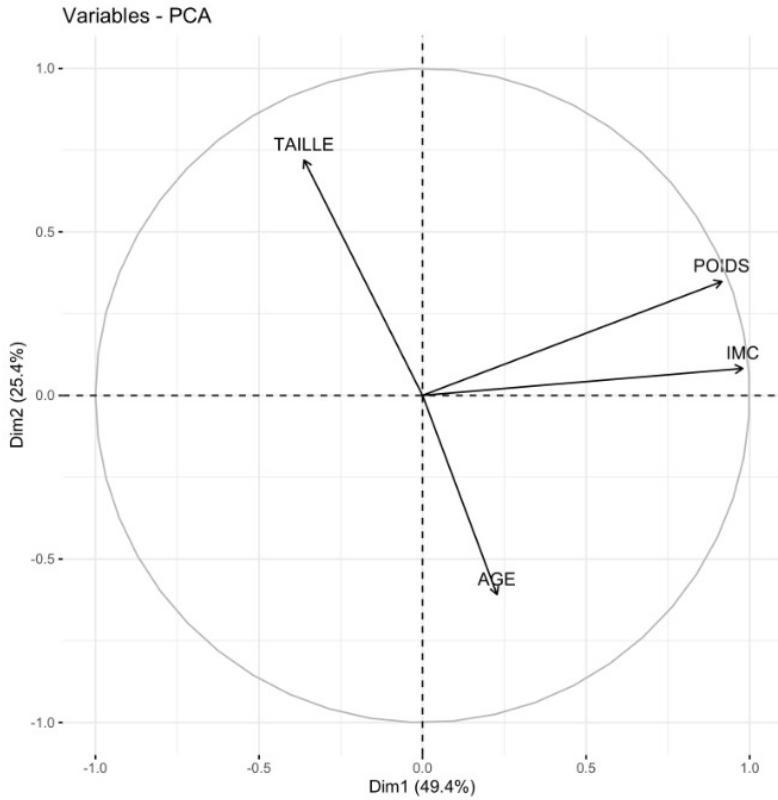


FIGURE 22 – Graphique de corrélation des variables : Cercle des corrélations

**Interprétation :** Les variables positivement corrélées sont regroupées (*poids, IMC*).

Les variables négativement corrélées sont positionnées sur les côtés opposés (quadrants opposés) de l'origine du graphique (*taille*). La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP.

Nous constatons que la *taille* est très proche du cercle et du premier axe principal (du côté des valeurs négatives) ce qui confirme sa forte corrélation négative avec la première composante principale. Elle est donc bien représentée par l'ACP et est en outre opposée à la variable *âge*. Les variables *poids* et *IMC* sont proches l'une de l'autre et du centre du cercle. Elles ne donnent donc aucune information sur leur corrélation avec les deux premières composantes principales. En somme, l'analyse du cercle des corrélations confirme les observations précédemment faites.

## 4.2 Analyse factorielle des correspondances multiples - ACM

L'Analyse des Correspondances Multiples est une extension de l'analyse factorielle des correspondances pour résumer et visualiser un tableau de données contenant plus de deux

variables qualitatives. L'objectif est d'identifier un groupe de personnes ayant un profil similaire dans leurs réponses aux questions et les associations entre les catégories des variables.

#### 4.2.1 Les valeurs propres

La proportion des variances retenues par les différentes dimensions (axes) peut être extraite, comme l'affiche la figure ci-dessous :

```
> eig.val
  eigenvalue variance.percent cumulative.variance.percent
Dim.1  0.3306747      26.45398                  26.45398
Dim.2  0.2739301      21.91441                  48.36839
Dim.3  0.2520320      20.16256                  68.53095
Dim.4  0.2061705      16.49364                  85.02460
Dim.5  0.1871925      14.97540                 100.00000
>
```

FIGURE 23 – Tableau des valeurs propres

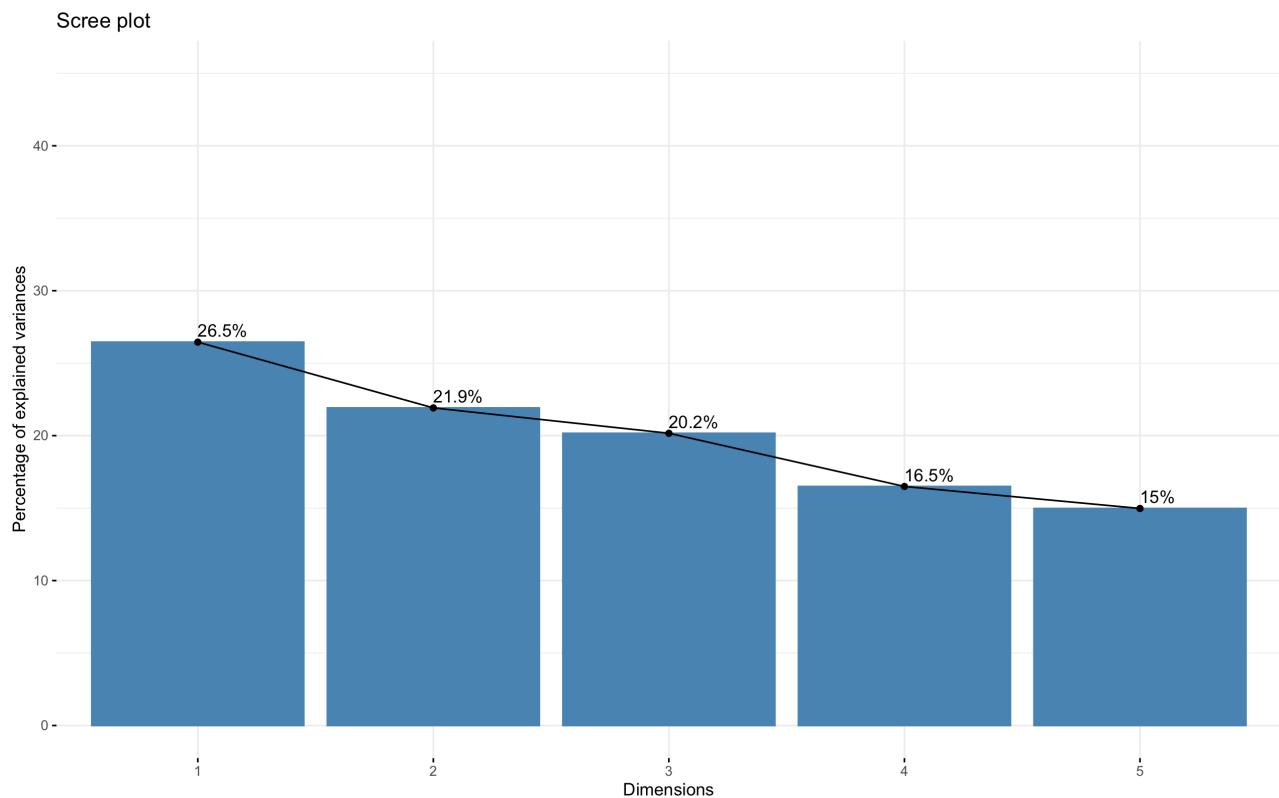


FIGURE 24 – Histogramme des valeurs propres

#### 4.2.2 Coordonnées des catégories des variables

Le graphique ci-dessous permet d'identifier les variables les plus corrélées avec chaque axe. Les corrélations au carré entre les variables et les axes sont utilisées comme coordonnées. On constate que la variable 'TABAC' est bien représentée par l'axe 1. Les couleurs changent du bleu au rouge proportionnellement à la valeurs des variables.

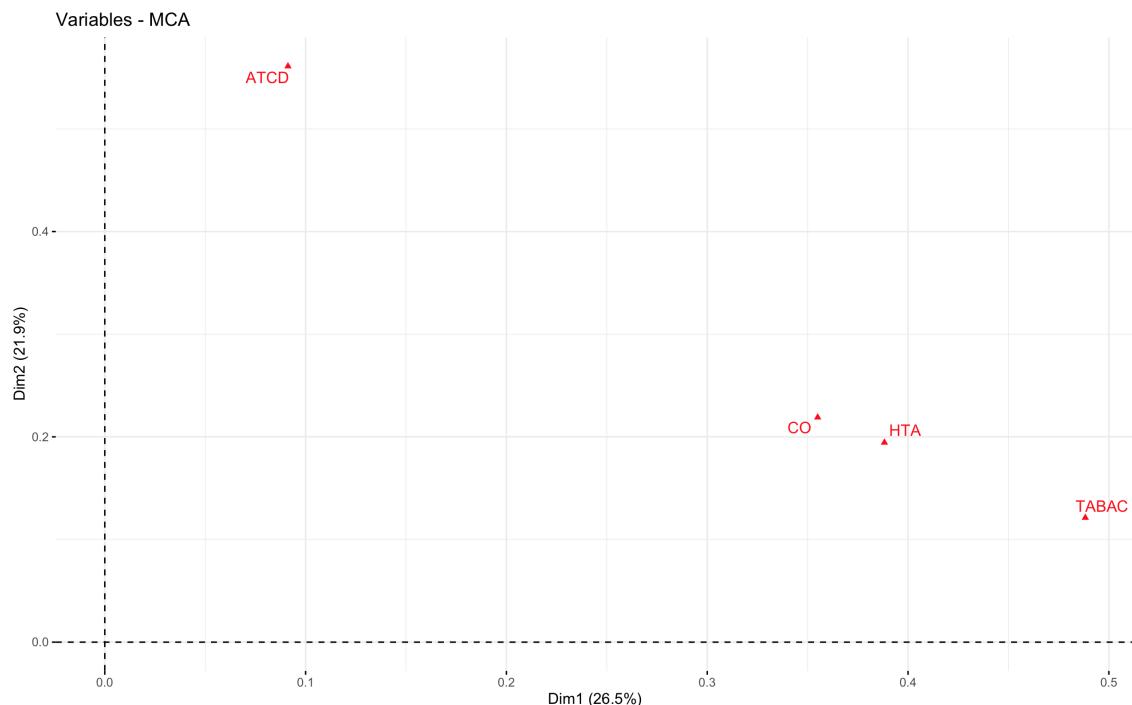


FIGURE 25 – Corrélation des variables par rapport aux axes ( dimensions)

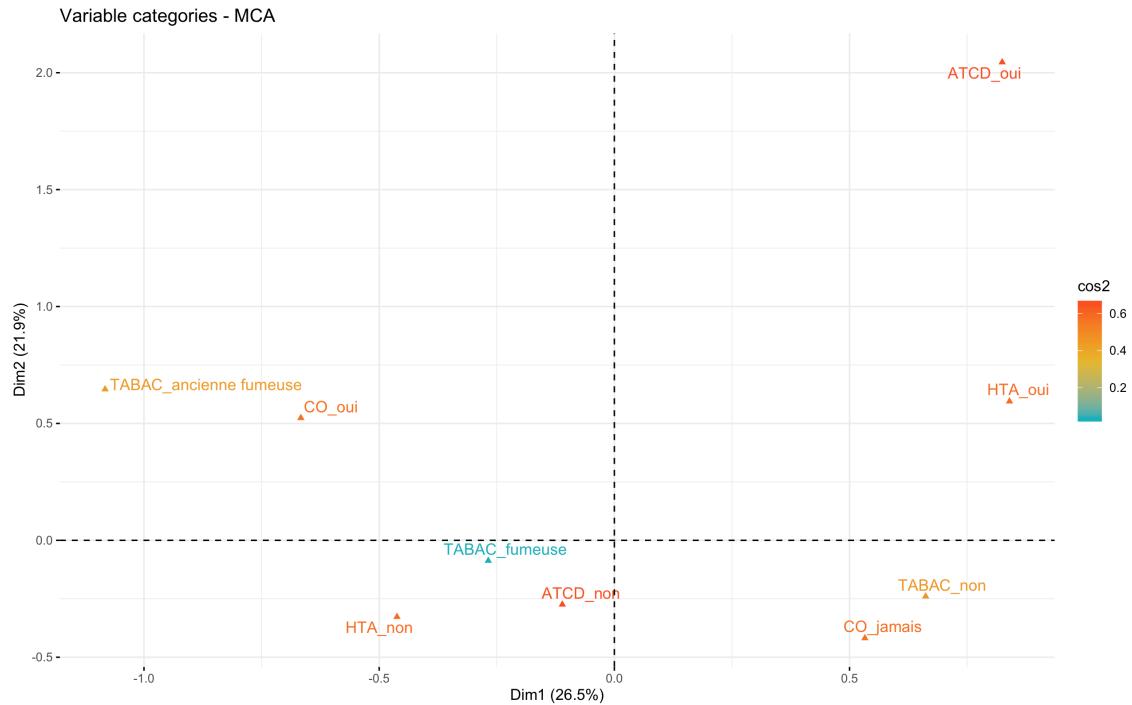


FIGURE 26 – Coordonnées des catégories des variables

Les catégories avec un profil similaire sont regroupées. Les catégories corrélées négativement sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés). La distance entre les catégories et l'origine mesure la qualité des catégories. **Les points qui sont loin de l'origine sont bien représentés par l'ACM.**

#### 4.2.3 Contribution des variables aux dimensions

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
CO_jamais	11.92	8.88	0.14	15.81	7.67
CO_oui	14.92	11.11	0.18	19.78	9.59
TABAC_ancienne fumeuse	19.39	8.34	18.72	9.91	21.76
TABAC_fumeuse	1.64	0.21	63.33	1.51	3.18
TABAC_non	15.89	2.52	11.46	1.33	20.82
ATCD_non	0.82	6.06	0.65	3.08	1.23
ATCD_oui	6.08	45.15	4.84	22.96	9.15
HTA_non	10.42	6.30	0.24	9.09	9.44
HTA_oui	18.93	11.45	0.44	16.53	17.16

FIGURE 27 – Contribution des variables aux dimensions

Les variables avec les plus grandes valeurs, contribuent le mieux à la définition des dimensions. Les catégories qui contribuent le plus à Dim.1 et Dim.2 sont les plus importantes pour expliquer la variabilité dans le jeu de données.

## Représentation graphique

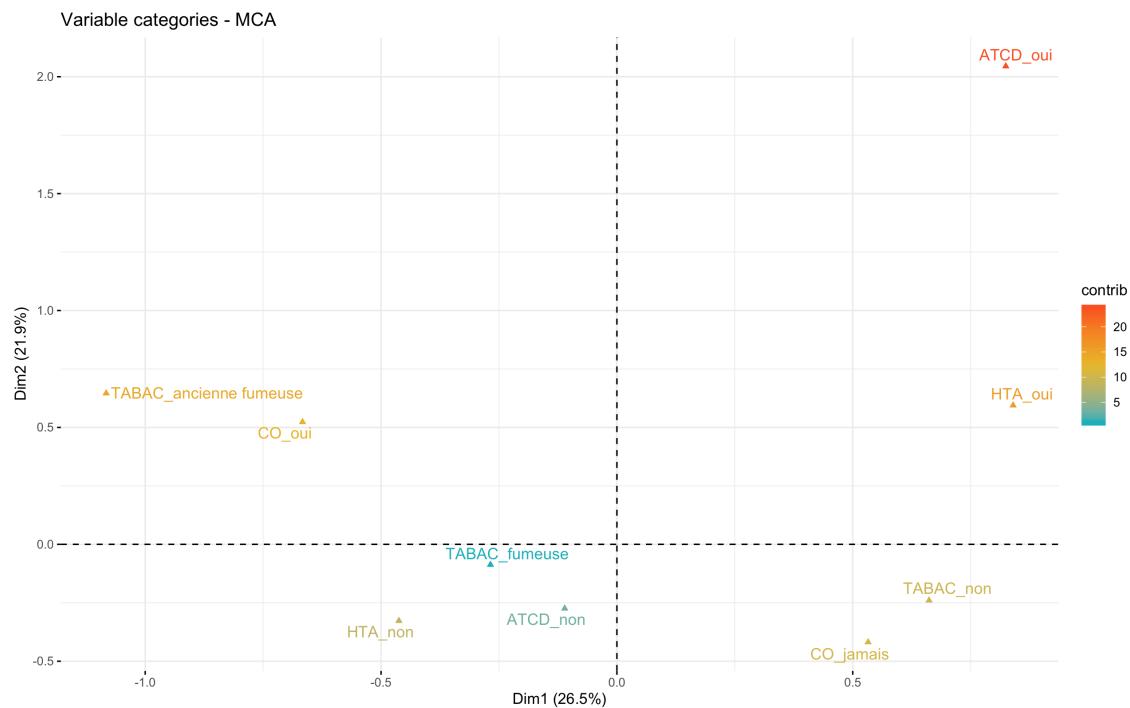


FIGURE 28 – Illustration - Contribution des variables aux dimensions

Le graphique ci-dessus donne une idée du pôle des dimensions auquel les catégories contribuent réellement. Les modalités **HTA\_oui** et **TABAC\_non** ont une contribution importante au pôle positif de la première dimension (axe). Tandis que les modalités **TABAC\_fumeuse** a une contribution minime au pôle négatif de la première dimension ; ainsi de suite.

#### 4.2.4 Spécification dans l'ACM

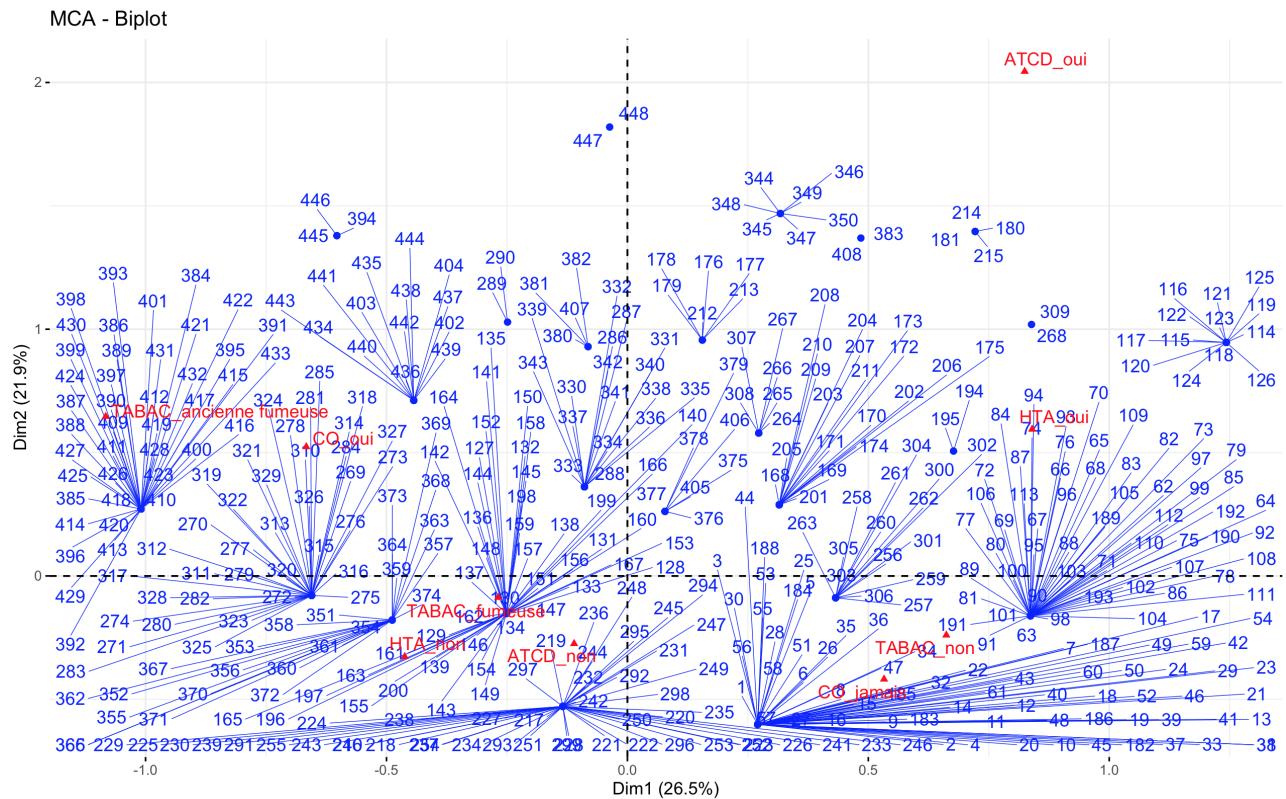


FIGURE 29 – Illustration dans l'ACM

Les individus actifs sont en bleu. Les modalités des variables actives sont en rouge.

### 4.3 Analyse factorielle des données mixtes- AFDM

La mise en oeuvre de la méthode AFDM est possible et simple avec **R** mais l'accès aux résultats est un peu plus compliqué en revanche. C'est pour cela que nous avons opté pour **Tanagra** pour réaliser notre AFDM.

#### 4.3.1 Le tableau des valeurs propres

## Eigen values

Matrix trace = 9.00

Axis	Eigen value	% explained	Histogram	% cumulated
1	2.135312	23.73%		23.73%
2	1.655303	18.39%		42.12%
3	1.152654	12.81%		54.93%
4	0.962170	10.69%		65.62%
5	0.925188	10.28%		75.90%
6	0.862626	9.58%		85.48%
7	0.692611	7.70%		93.18%
8	0.557151	6.19%		99.37%
9	0.056986	0.63%		100.00%
Tot.	9.000000	-	-	-

FIGURE 30 – Tableau des valeurs propres

En calculant le seuil ( $1 + 1,65 * \text{racineCarré}(p - 1/n - 1)$ ), où ‘p’ est le nombre total de valeurs propres théoriquement non nulles ( $p = 9$ ), et ‘n’ le nombre d’observations ( $n = 448$ ), le seuil est égal à **1.22**. Ceci nous conduit à ne considérer que les deux premiers axes (*comme l’indique Tanagra sur la figure ci-dessus*) mais le pourcentage indiqué (**42.12%**) est trop peu. Nous risquons d’ignorer des informations importantes. Du coup, nous nous sommes proposés de prendre jusqu’à **75.90%** pour ne laisser tomber aucune information. Les cinq (05) premiers axes traduisent donc 75.90% de l’information disponible.

### 4.3.2 Le tableau des coordonnées

Le tableau des coordonnées décrit l'impact des variables, qu'elles soient quantitatives ou qualitatives, dans la définition des axes. Les variables marquées (\*) indiquent le carré du coefficient de corrélation linéaire ; pour (\*\*), les valeurs correspondent au carré du rapport de corrélation.

**Squared Correlation (Communalities)**

Attribute	Axis_1			Axis_2			Axis_3			Axis_4			Axis_5		
	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)
CO (**)	0.006876	0.3 %	1 % (1 %)	0.306075	18.5 %	31 % (31 %)	0.006483	0.6 %	1 % (32 %)	0.086116	9.0 %	9 % (41 %)	0.180924	19.6 %	18 % (59 %)
TABAC (**)	0.049653	2.3 %	2 % (2 %)	0.511535	30.9 %	26 % (28 %)	0.138025	12.0 %	7 % (35 %)	0.685764	71.3 %	34 % (69 %)	0.153175	16.6 %	8 % (77 %)
AGE (*)	0.100787	4.7 %	10 % (10 %)	0.504505	30.5 %	50 % (61 %)	0.010143	0.9 %	1 % (62 %)	0.004179	0.4 %	0 % (62 %)	0.000750	0.1 %	0 % (62 %)
POIDS (*)	0.781877	36.6 %	78 % (78 %)	0.049028	3.0 %	5 % (83 %)	0.018329	1.6 %	2 % (85 %)	0.024028	2.5 %	2 % (87 %)	0.054363	5.9 %	5 % (93 %)
TAILLE (*)	0.089206	4.2 %	9 % (9 %)	0.003893	0.2 %	0 % (9 %)	0.469491	40.7 %	47 % (56 %)	0.001114	0.1 %	0 % (56 %)	0.354249	38.3 %	35 % (92 %)
IMC (*)	0.864085	40.5 %	86 % (86 %)	0.055169	3.3 %	6 % (92 %)	0.010815	0.9 %	1 % (93 %)	0.028131	2.9 %	3 % (96 %)	0.000476	0.1 %	0 % (96 %)
ATCD (**)	0.060532	2.8 %	6 % (6 %)	0.000026	0.0 %	0 % (6 %)	0.399046	34.6 %	40 % (46 %)	0.072369	7.5 %	7 % (53 %)	0.159822	17.3 %	16 % (69 %)
HTA (**)	0.182296	8.5 %	18 % (18 %)	0.225072	13.6 %	23 % (41 %)	0.100321	8.7 %	10 % (51 %)	0.060470	6.3 %	6 % (57 %)	0.021428	2.3 %	2 % (59 %)
Var. Expl.	2.135312	-	24 % (24 %)	1.655303	-	18 % (42 %)	1.152654	-	13 % (55 %)	0.962170	-	11 % (66 %)	0.925188	-	10 % (76 %)

(\*) Square of correlation coefficient

(\*\*) Correlation ratio

FIGURE 31 – Tableau des coordonnées

Le pourcentage en ligne (%) indique la part d'information de la variable qui est retranscrite par l'axe. Pour la variable **CO**, 31% de l'information qu'elle véhicule est restituée par le second axe. De même que les variables **tabac** et **age** dont les pourcentages sont respectivement 28% et 61% au total.

Les variables **poids** et **IMC** sont celles qui pèsent plus sur la définition du premier axe de la façon dont **CO**, **tabac**, **age** et **HTA** pèsent sur le second axe.

### 4.3.3 Le tableau des corrélations

Ici, nous verrons la relation entre les variables.

## Continuous Attributes - Correlation (Factor Loadings)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4	Axis_5
AGE	-0.317470	-0.710285	0.100714	0.064643	0.027387
POIDS	-0.884238	0.221422	-0.135383	0.155009	-0.233159
TAILLE	0.298674	-0.062395	-0.685194	-0.033378	-0.595188
IMC	-0.929562	0.234880	0.103994	0.167723	-0.021820

FIGURE 32 – Tableau des corrélations

Sur le premier axe, nous remarquons que le **poids** et l'**IMC** sont fortement liées mais de façon négative. Les individus avec un poids élevé ont également un indice de masse corporelle qui s'accompagne. Sur le second axe, l'**age** s'oppose au **poids** et à l'**IMC**.

### 4.3.4 Le tableau des moyennes conditionnelles

Ce tableau positionne les modalités sur les axes factoriels. Nous avons également une indication sur leurs contributions. Elles dépendent à la fois de l'écartement avec l'origine, de l'effectif et de la valeur propre associée à l'axe.

#### Discrete Attributes - Conditional means and contributions

Attribute	Axis_1			Axis_2			Axis_3			Axis_4			Axis_5			
	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	
CO	jamais	-0.1083	0.14	-1.753	-0.6363	8.21	-11.697	0.0773	0.25	1.702	0.2573	3.98	6.204	-0.3658	8.69	-8.993
	oui	0.1355	0.18	1.753	0.7962	10.28	11.697	-0.0967	0.31	-1.702	-0.3220	4.97	-6.204	0.4577	10.87	8.993
	Tot.	-	0.32	-	-	18.49	-	-	0.56	-	-	8.95	-	-	19.56	-
TABAC	non	-0.1046	0.12	-1.453	-0.7997	11.20	-12.623	0.3168	3.62	5.992	-0.4826	12.08	-9.993	-0.0474	0.13	-1.000
	fumeuse	0.4624	1.41	4.394	0.1393	0.21	1.504	-0.5970	8.08	-7.721	1.2334	49.51	17.459	0.4852	8.29	7.004
	ancienne fumeuse	-0.4076	0.80	-3.121	1.5624	19.49	13.586	0.1274	0.27	1.328	-0.6402	9.68	-7.301	-0.5645	8.14	-6.564
	Tot.	-	2.33	-	-	30.90	-	-	11.97	-	-	71.27	-	-	16.56	-
ATCD	non	0.1317	0.34	5.202	0.0024	0.00	0.109	0.2484	4.10	13.356	0.0967	0.89	5.688	-0.1409	2.04	-8.452
	oui	-0.9815	2.50	-5.202	-0.0180	0.00	-0.109	-1.8515	30.52	-13.356	-0.7204	6.63	-5.688	1.0498	15.23	8.452
	Tot.	-	2.83	-	-	0.00	-	-	34.62	-	-	7.52	-	-	17.27	-
HTA	non	0.4628	3.03	9.027	0.4527	4.83	10.030	0.2522	3.09	6.697	0.1789	2.23	5.199	-0.1044	0.82	-3.095
	oui	-0.8411	5.51	-9.027	-0.8229	8.77	-10.030	-0.4585	5.61	-6.697	-0.3252	4.05	-5.199	0.1898	1.49	3.095
	Tot.	-	8.54	-	-	13.60	-	-	8.70	-	-	6.28	-	-	2.32	-

FIGURE 33 – Tableau des moyennes conditionnelles

Prenons le cas du **tabac** sur le second axe, le carré du rapport de corrélation **0.51**, sa contribution est donc de **30.90%**; avec une forte participation des non fumeuses (**11.2%**) et des anciennes fumeuses (**19.49%**).

#### 4.3.5 Représentation graphique des individus

##### — Projection des individus

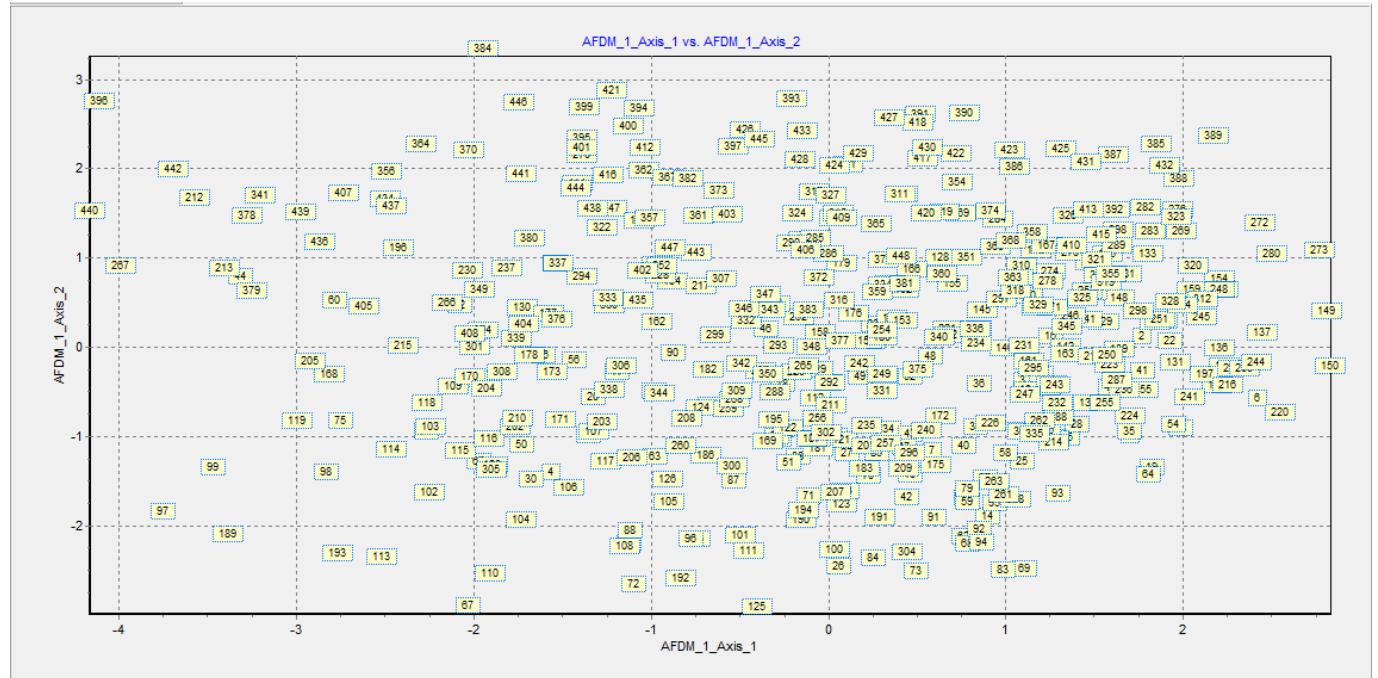


FIGURE 34 – La projection des individus

Nous représentons ici, la projection des individus par rapport à l'axe 1 et 2, avec comme attribut de référence **INFARCTUS**.

#### 4.3.6 Le cercle de corrélation

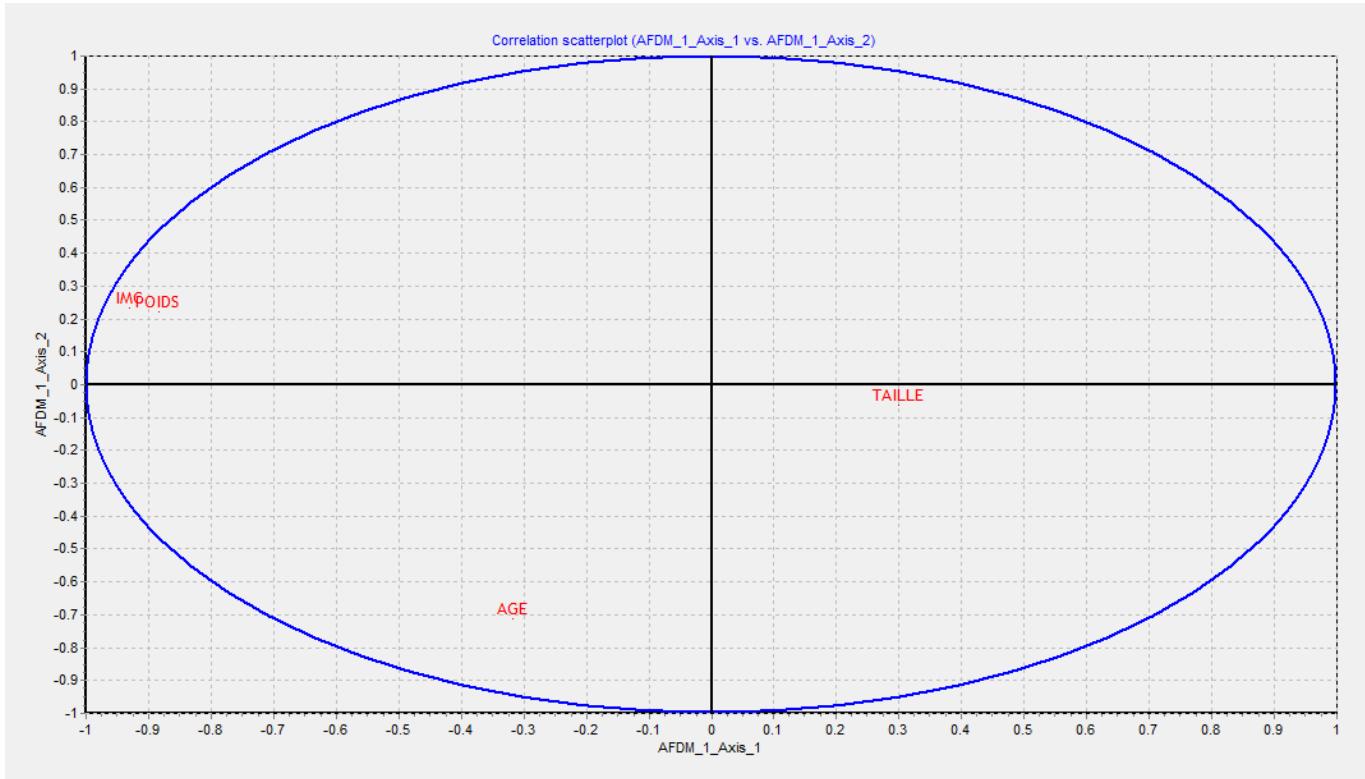


FIGURE 35 – Le cercle de corrélation

Ce graphique illustre ce qui a été dit plus haut, au niveau du tableau de corrélation.

#### 4.3.7 Représentation des moyennes conditionnelles

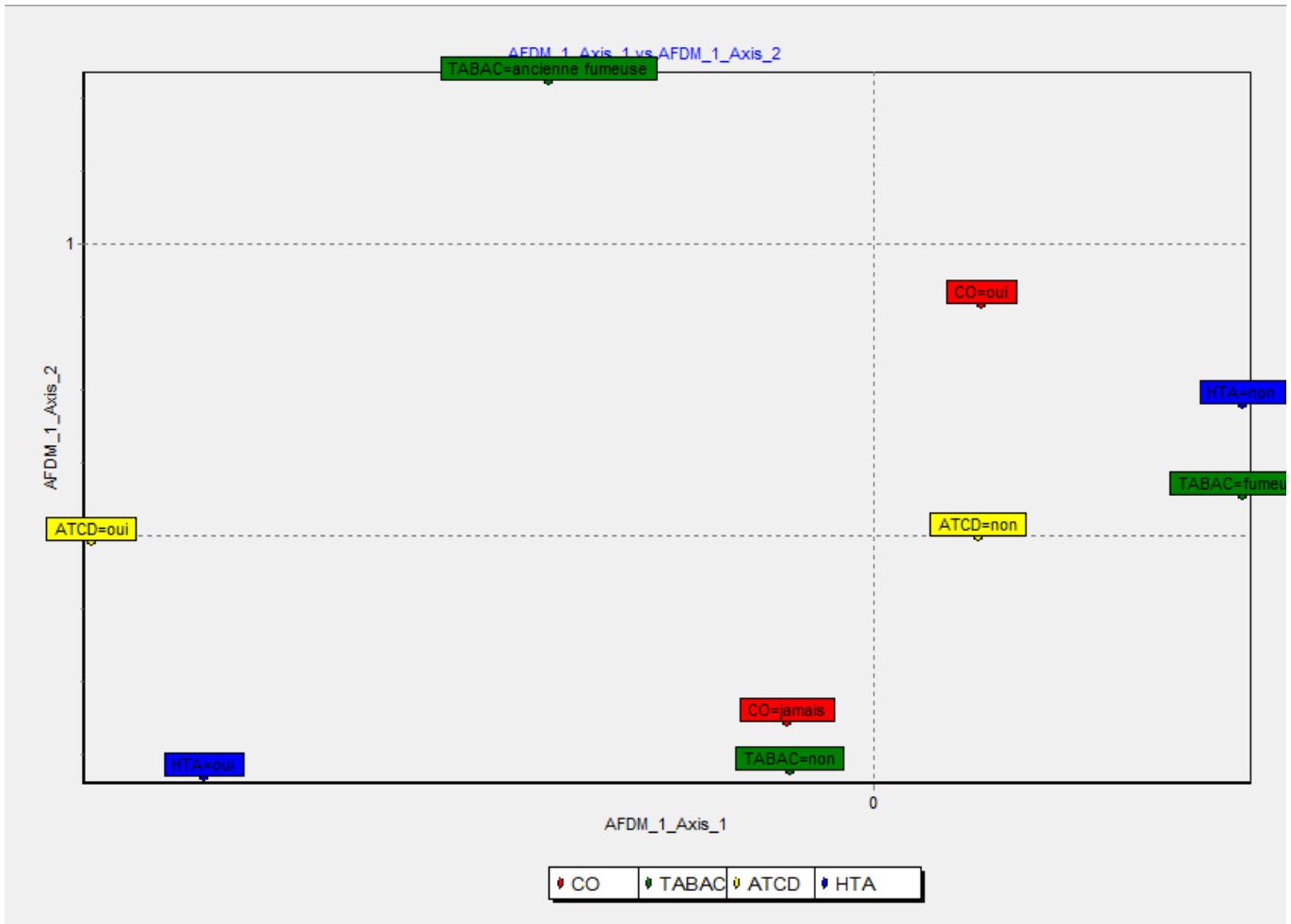


FIGURE 36 – Illustration des moyennes conditionnelles

Pour renforcer l'interprétation faite ci-dessus, la modalité '**TABAC=ancienne fumeuse**' pèse dans notre analyse.

## 5 Classification - Clustering

Les méthodes de classification sont utilisées pour analyser des données multivariées. L'objectif principal consiste à soit identifier des groupes d'individus présentant des traits communs ou de partitionner les individus en plusieurs groupes sur la base de traits communs. Étant donné que nous disposons de données mixtes de variables continues et catégorielles, nous allons appliquer la classification sur les résultats de l'AFDM réalisée plus haut.

## 5.1 La classification Ascendante hiérarchique

Avec **R**, la fonction qui permet de faire la classification des données s'appelle **HCPC()** pour *Hierarchical Clustering on Principal Components*. Elle comporte les méthodes de classification (CAH et k-Means) et les méthodes de composantes principales (ACP, AFC, AFM, etc). En d'autres termes, elle applique les méthodes de classification sur les résultats des analyses factorielles (ACP, ACM, AFM, etc).

Le HCPC peut être utile des situations telles que lorsqu'on dispose d'un grand nombre de variables continues dans votre jeu de données et le Clustering sur des variables catégorielles. Et vu que nous disposons de données mixtes, de variables continues et catégorielles, nous allons effectuer la classification sur les résultats de l'AFDM réalisé plus haut.

### 5.1.1 Le dendrogramme

Le dendrogramme est un diagramme fréquemment utilisé pour illustrer l'arrangement de groupes générés par un regroupement hiérarchique ou hiérarchisant. Celui lié à notre analyse de présente comme suit :

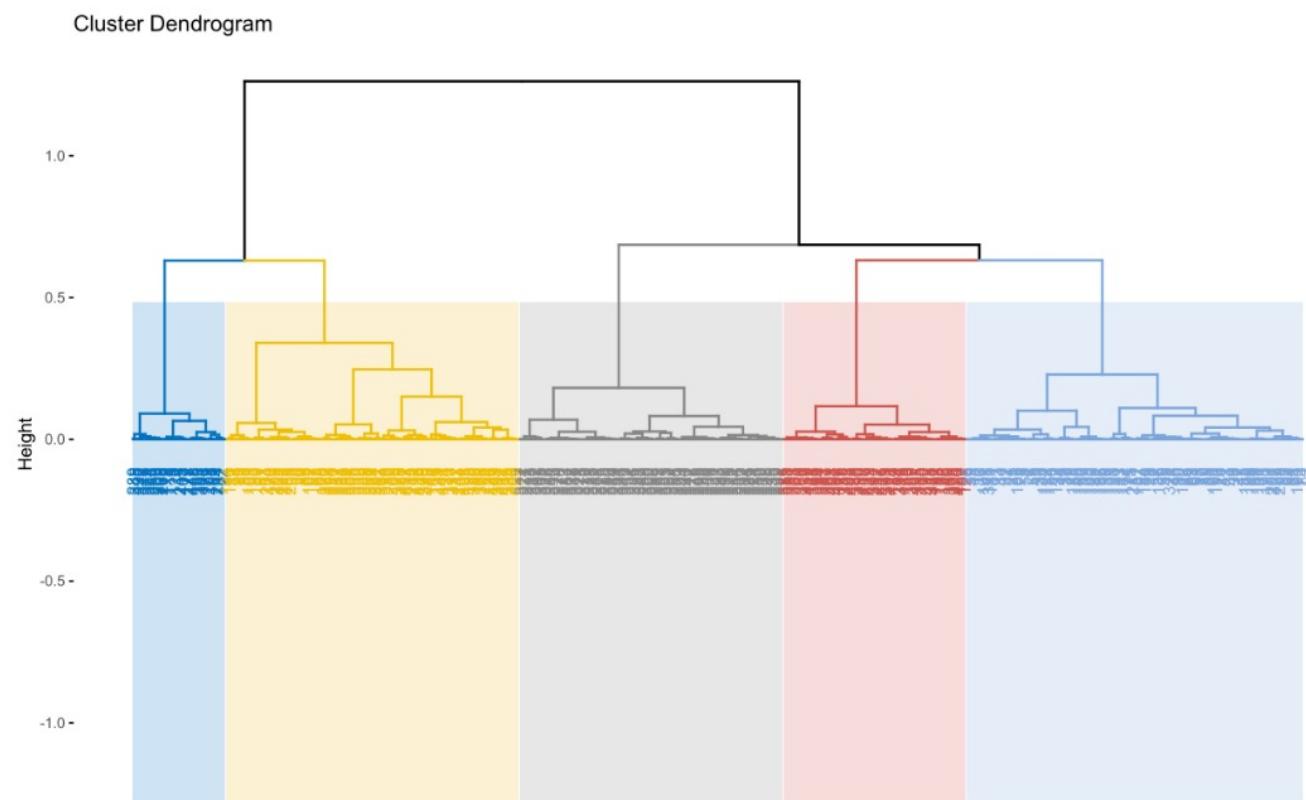


FIGURE 37 – Le dendrogramme issu de notre classification

Le dendrogramme nous suggère une solution à 5 groupes.

Le plan facteur qui s'en suit :

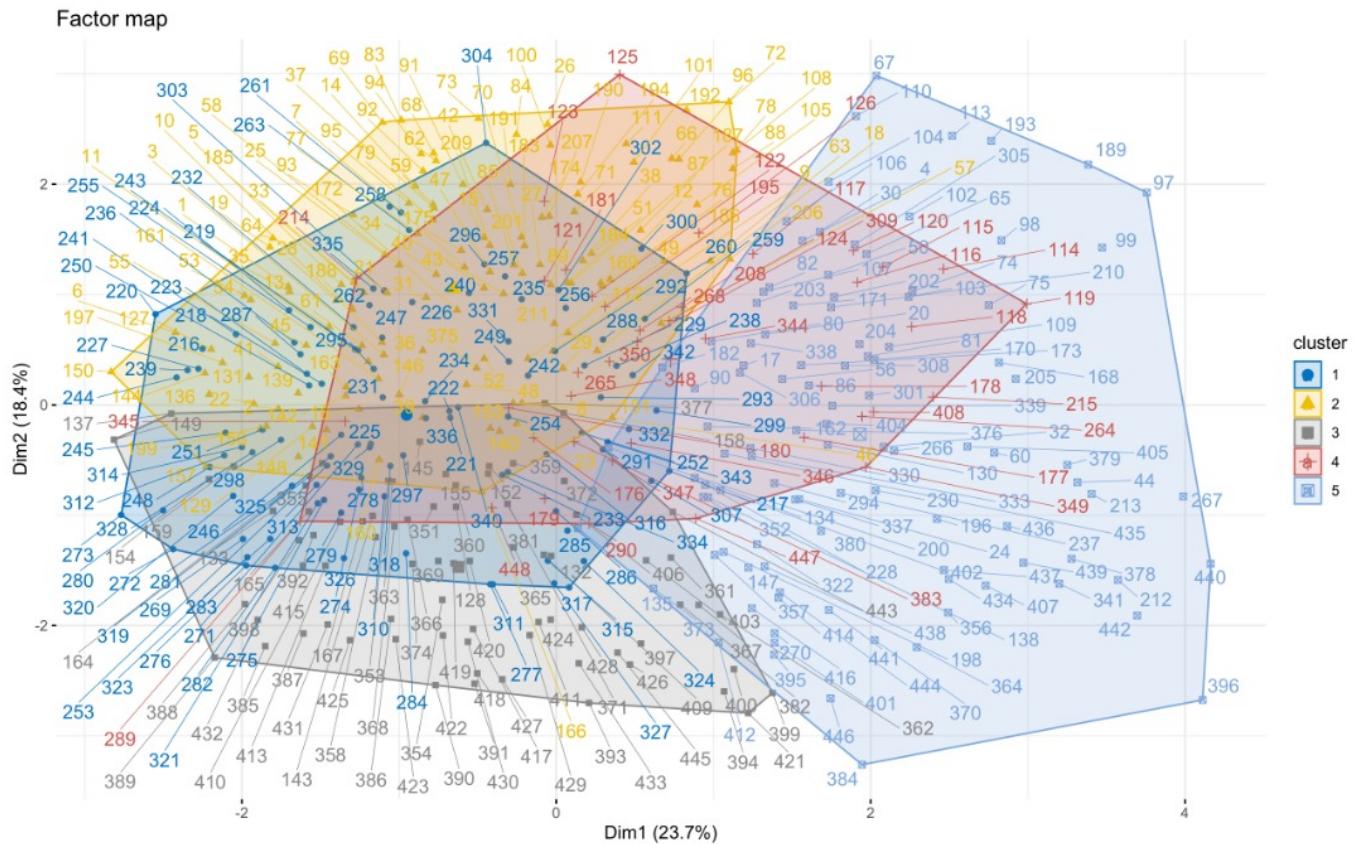


FIGURE 38 – Le plan facteur issu de notre classification

### 5.1.2 Le Graphique 3D combinant la classification hiérarchique et le plan des facteurs

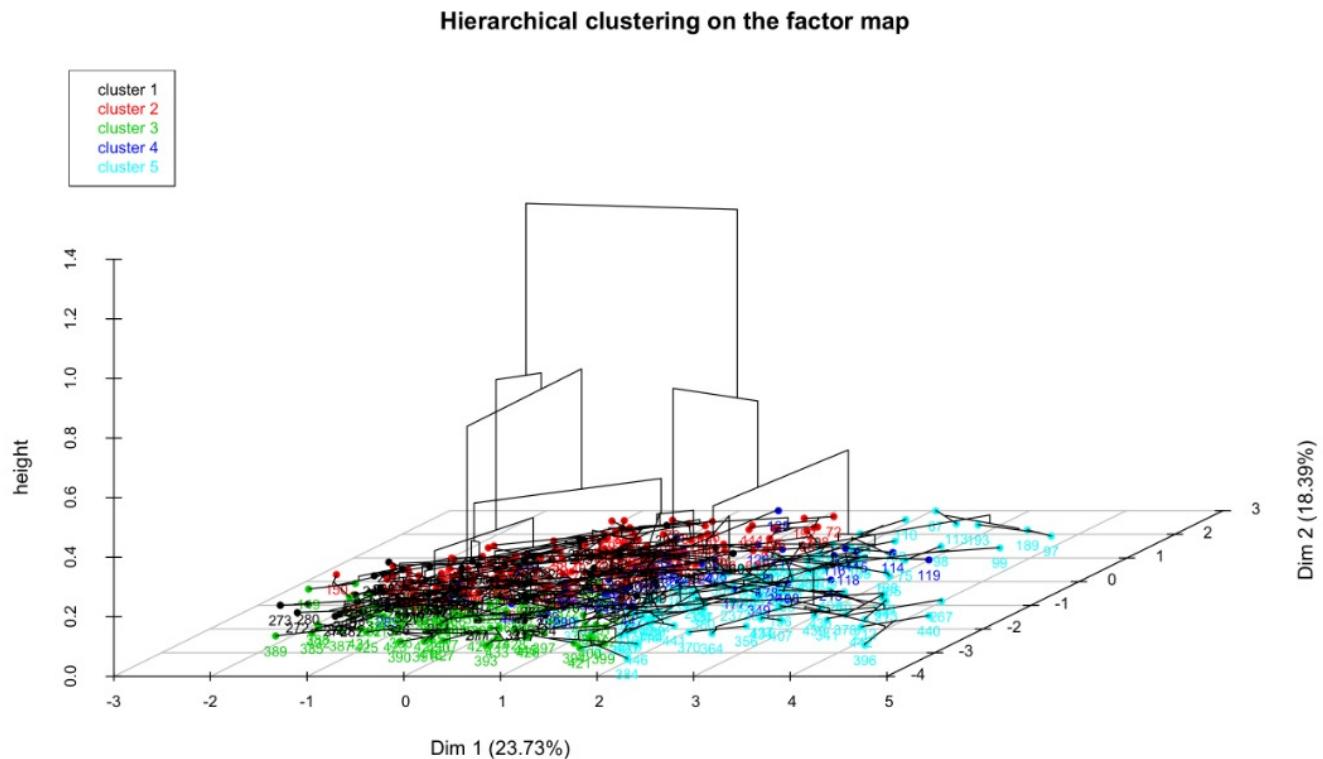


FIGURE 39 – Le Graphique 3D : classification hiérarchique & plan des facteurs

### 5.1.3 Variables quantitatives décrivant le plus chaque cluster

v.test Mean in category Overall mean sd in category Overall sd p.value						
IMC	-4.986378	21.14193	24.20312	4.347823	7.136682	6.152168e-07
POIDS	-5.090027	58.54808	66.44866	12.380102	18.043896	3.580128e-07
\$`2`						
v.test Mean in category Overall mean sd in category Overall sd p.value						
AGE	4.676942	51.36000	45.61161	15.941092	16.165606	2.911851e-06
POIDS	-7.639516	55.96800	66.44866	11.681052	18.043896	2.180396e-14
IMC	-8.089508	19.81366	24.20312	3.807469	7.136682	5.990624e-16
\$`3`						
v.test Mean in category Overall mean sd in category Overall sd p.value						
TAILLE	3.889552	168.44737	165.14955	6.632049	8.102438	1.004293e-04
IMC	-2.657909	22.21818	24.20312	4.556826	7.136682	7.862710e-03
AGE	-7.642020	32.68421	45.61161	10.254922	16.165606	2.138402e-14
\$`4`						
v.test Mean in category Overall mean sd in category Overall sd p.value						
TAILLE	2.676138	168.425	165.1496	7.816289	8.102438	0.007447607
\$`5`						
v.test Mean in category Overall mean sd in category Overall sd p.value						
IMC	16.014205	34.09637	24.20312	4.857667	7.136682	1.016980e-57
POIDS	13.629876	87.73786	66.44866	13.996505	18.043896	2.660302e-42
AGE	3.267824	50.18447	45.61161	17.462163	16.165606	1.083778e-03
TAILLE	-7.701940	159.74757	165.14955	7.885982	8.102438	1.340160e-14

FIGURE 40 – Variables quantitatives décrivant le plus chaque cluster

Les variables IMC et POIDS sont associées au cluster 1 mais de façon négative.

La variable AGE est la plus significativement associée au cluster 2. Les variables POIDS et IMC, toujours de façon négative.

La variable TAILLE est la plus significativement associée au cluster 3. Les variables AGE et IMC y sont associées de façon négative.

Dans le cluster 4, la variable TAILLE est la seule et la plus associée.

La variable IMC, POIDS, AGE sont plus associées au cluster 5. La variable TAILLE l'est aussi mais de façon négative. Par exemple, la valeur moyenne de la variable TAILLE dans le cluster 5 est de **159.74757**, ce qui est inférieure à la moyenne globale (**165.14955**) dans tous les clusters. Par conséquent, on peut conclure que le cluster 5 se caractérise par un faible taux de la variable TAILLE par rapport à tous les autres clusters où la TAILLE est représentée.

### 5.1.4 Axes principaux associés aux clusters

```
> res.hcpc$desc.axes$quanti
$`1`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.4 15.280164      1.2893260 1.320444e-15      0.4286465 0.9809028 1.036797e-52
Dim.3  3.692167      0.3409883 3.830312e-16      0.6875366 1.0736170 2.223514e-04
Dim.5 -4.226141     -0.3496775 -3.304764e-16      0.7886080 0.9618668 2.377333e-05
Dim.1 -7.576328     -0.9523536 -2.489948e-16      0.9005872 1.4612707 3.554716e-14

$`2`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.2 10.777164      1.0542308 9.393450e-16      0.8492826 1.2865856 4.412814e-27
Dim.5  3.099613      0.2266809 -3.304764e-16      0.8387926 0.9618668 1.937737e-03
Dim.1 -5.889052     -0.6542874 -2.489948e-16      0.8927716 1.4612707 3.884176e-09
Dim.4 -6.281446     -0.4684660 1.320444e-15      0.4110454 0.9809028 3.354390e-10
Dim.3 -6.447968     -0.5263380 3.830312e-16      0.6814592 1.0736170 1.133598e-10

$`3`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.5   6.665477      0.6708988 -3.304764e-16      0.7981869 0.9618668 2.638077e-11
Dim.1  -4.068960     -0.6221929 -2.489948e-16      0.9409158 1.4612707 4.722341e-05
Dim.4  -7.014258     -0.7199770 1.320444e-15      0.3693976 0.9809028 2.311720e-12
Dim.2 -10.907228    -1.4684663 9.393450e-16      0.7235904 1.2865856 1.064534e-27

$`4`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.3 13.050914      2.1165895 3.830312e-16      0.7933812 1.0736170 6.278990e-39
Dim.1   3.300791      0.7286098 -2.489948e-16      1.0757623 1.4612707 9.641274e-04
Dim.2   2.137627      0.4154483 9.393450e-16      0.8936013 1.2865856 3.254703e-02
Dim.4  -6.056918     -0.8974783 1.320444e-15      0.9113106 0.9809028 1.387543e-09
Dim.5  -7.283308     -1.0582539 -3.304764e-16      0.8093855 0.9618668 3.257333e-13

$`5`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.1 15.271757      1.9317764 -2.489948e-16      0.8601187 1.461271 1.179516e-52
Dim.2  -2.384617     -0.2655794 9.393450e-16      1.3313395 1.286586 1.709692e-02
Dim.3  -5.133786     -0.4771164 3.830312e-16      0.9801335 1.073617 2.839703e-07
```

FIGURE 41 – Axes principaux associés aux clusters

Les résultats ci-dessus indiquent que :

- les individus dans les groupes 1 ont des coordonnées élevées sur l'axe 4, 3, 5 et 1.
- les individus du groupe 2 ont des coordonnées élevées sur tous les axes mais surtout l'axe 2
- les individus appartenant au groupe 3 ont des coordonnées élevées sur les axes 5, 1, 4 et 2
- les individus du groupe 4 ont des coordonnées élevées sur tous les axes mais surtout l'axe 3

- les individus du groupe 5 ont des coordonnées élevées sur les axes 1, 2 et 3.

## 5.2 La méthode de K-MEANS

Cette méthode subdivise les individus en k-groupes, k étant le nombre optimal de groupes à définir. Elle est également utilisée pour caractériser les individus d'une même classe. Nous allons dans un premier temps déterminer le nombre **k** de clusters avant de passer à l'application de l'algorithme de K-Means proprement dite.

### 5.2.1 Estimation du nombre optimal de clusters

Il existe dans **R** des fonctions permettant de déterminer le nombre de clusters à utiliser pour une analyse. Il s'agit principalement de **fviz\_nbclust()** du package **Factoextra** de R. L'opération faite est la suivante :

- convertir en numérique les variables qualitatives
- scale le nouveau dataset : `datasetNumeric.scaled <- scale(datasetNumeric)`
- `fviz_nbclust(datasetNumeric.scaled, kmeans, method = "silhouette")`

Ce qui nous donne le résultat suivant :

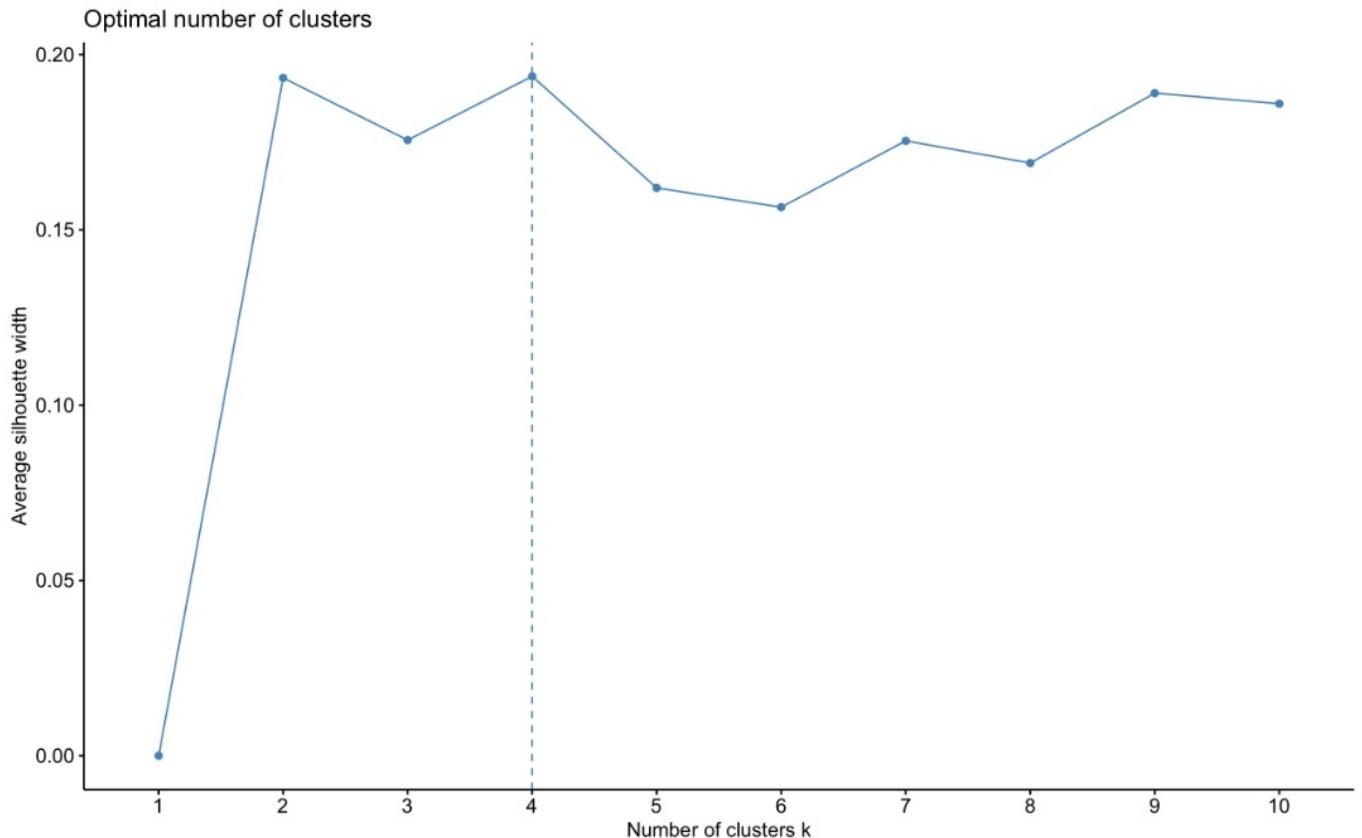


FIGURE 42 – Détermination du nombre k de clusters

Le graphique ci-dessus représente la variance au sein des clusters. Il diminue à mesure que k augmente, mais on peut voir un pli à  $k = 4$ . Ce pli indique que les clusters supplémentaires au-delà du quatrième ont peu de valeur. Dans la section suivante, nous allons classer les observations en **4 clusters**.

### 5.2.2 Classement des observations

Pour la réalisation, nous avons fait :

— `km <- kmeans(datasetNumeric.scaled, 4, nstart = 25)`

Ce qui nous donne le résultat suivant :

Pour la visualisation graphique : `fviz_cluster(km, data = datasetNumeric.scaled )`. Ce qui nous donne :

Cluster plot

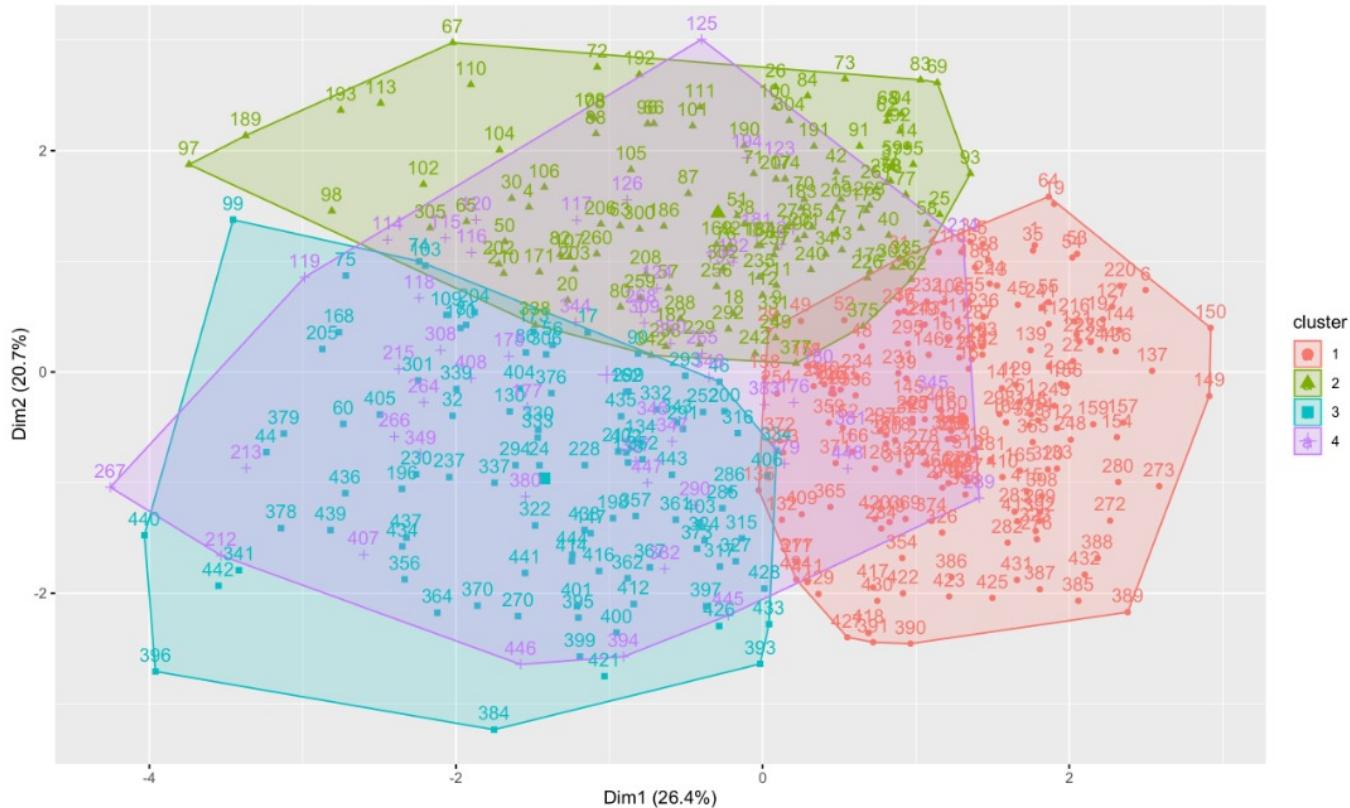


FIGURE 43 – Visualisation graphique de la méthode K-Means

# Conclusion

Pour ce travail de recherche, nous avons choisi le domaine médical surtout les données sur l'infarctus. Grâce au logiciel R et par moment du logiciel Tanagra, nous avons pu faire les études statistiques sur les variables qualitatives et quantitatives du jeu de données choisi. Nous avons rencontrés des problèmes concernant la forme des variables et les résultats des traitements mais les recherches et les aides de la part du professeur nous ont aidé à parvenir à arriver à bout de ce travail.

## Références

- Source de données : <http://www.biostatisticien.eu/springeR/jeuxDonnees5.html>
- <https://support.minitab.com/fr-fr/minitab/18/help-and-how-to/statistics/tables/how-to/chisquare-test-for-association/interpret-the-results/key-results/>
- <https://sites.google.com/site/rgraphiques/home>
- <http://rstudio-pubs-static.s3.amazonaws.com/52670156db47a0604aa9818143e7d2db226e.html>
- <http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/78-classification-hierarchique-sur-composantes-principales-l-essentiel/>
- <http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/78-classification-hierarchique-sur-composantes-principales-l-essentiel/>
- <http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/87-cah-classification-ascendante-hierarchique-dans-r-avec-factominer-cours/>
- <https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/>