Université de Reims
Champagne-Ardenne

Rapport de projet :
AFDM avec R pour la prévention des accidents vasculaires cérébraux: Identification des facteurs de risques.

Etudier les similitudes entre les individus en prenant en compte des variables mixtes et d'étudier les relations entre toutes les variables (variables qualitatives et quantitatives).

JANVIER 2023

RÉALISÉ PAR

Khaireddine SATOURI
Idir HAREB
Youssouf Anis DAHLOUK
Diyane David NONON SAA

Table des matières

Table des matières 0
Introduction 1
Abstrait 1
Choix de la base de données 1
Comprenons bien le problème : 2
Choix de la méthode d'analyse 3
Informations sur le jeu de données et liste des variables 4
Processus d'analyse de l'AFDM: 4
AFDM avec FactoMineR 9
Conclusion 25
Les difficultés rencontrées 26
Bibliographie 26

Table de figures

Figure 1 : Méthodes d'Analyse 3
Figure 2 : Graphe des individus 10
Figure 3 : Graphe des variables 11
Figure 4 : Graphe des modalités 12
Figure 5 : Cercle de corrélation des variables quantitatives 12
Figure 6 : variable AFMD 13
Figure 7 : Contribution des variables à la première Dimension 14
Figure 8 : Contribution des variables à la Deuxième Dimension 15
Figure 9 : Cercle de corrélation des variables quantitatives avec contribution 16
Figure 10 : Modalités des variables qualitatives avec contribution 17
Figure 11 : Status d'AVC 18
Figure 12 : Statut tabagique 19
Figure 13 : Statut d'état matrimonial 20
Figure 14 : Type de profession 21
Figure 15 : Statut des maladies cardiaques 23
Figure 16 : Status de l'hypertension 24

Introduction

L'analyse des données (aussi appelée analyse exploratoire des données ou AED) est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives, elle permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci.

Elle comprend: l'analyse en composantes principales (ACP), employée pour des données quantitatives, et ses méthodes dérivées, l'analyse factorielle des correspondances (AFC) utilisée sur des données qualitatives, l'analyse factorielle des correspondances multiples (AFCM) généralisant la précédente et L'Analyse factorielle de données mixtes (AFDM) qui est une méthode factorielle dédiée aux tableaux dans lesquels un ensemble d'individus est décrit par un ensemble de variables quantitatives et qualitatives.

Abstrait

L'objectif de ce projet est d'utiliser l'analyse factorielle pour identifier la combinaison des caractéristiques qui sont plus susceptibles d'être associées à un accident vasculaire cérébral (AVC). Pour cette analyse, nous avons d'abord effectué une analyse exploratoire des données et une ingénierie des caractéristiques. Nous avons ensuite utilisé une technique de réduction de dimension adaptée aux ensembles de données mixtes de variables continues et catégorielles avec le package AFDM. Cette visualisation nous a permis d'identifier les caractéristiques les plus probablement associées au risque de développer un AVC.

Choix de la base de données

Selon l'Organisation mondiale de la santé (OMS), l'accident vasculaire cérébral (AVC) est la deuxième cause de décès dans le monde, responsable d'environ 11 % du total des décès. Il s'agit d'un autre problème de santé qui est en augmentation dans le monde entier en raison de l'adoption de changements de mode de vie qui ne tiennent pas compte d'un mode de vie sain et de bonnes habitudes alimentaires. Ainsi, les nouveaux appareils électroniques émergents qui enregistrent les signes vitaux de la santé ont ouvert la voie à la création d'une solution automatisée reposant sur des techniques d'IA. Ainsi, à l'instar des maladies cardiaques, des efforts ont été entrepris pour créer des tests de laboratoire permettant de prédire les accidents vasculaires cérébraux.

De nombreux ensembles de données qu'un scientifique des données rencontrera dans le monde réel contiendront à la fois des variables numériques et catégorielles. L'analyse factorielle des données mixtes (AFDM) est une méthode en composantes principales qui combine l'analyse en composantes principales (ACP) pour les variables continues l'analyse

des correspondances multiples (ACM) pour les variables catégorielles. C'est pourquoi nous avons choisi d'utiliser cette technique.

L'ensemble de données présenté ici comporte de nombreux facteurs (comme le sexe, l'âge, l'hypertension, les maladies cardiaques etc.) qui mettent en évidence le mode de vie des patients et nous donne donc l'occasion de créer une solution basée sur l'AFDM.

Comprenons bien le problème :

Un accident vasculaire cérébral se produit lorsqu'une partie du cerveau perd son approvisionnement en sang et cesse de fonctionner. Cela entraîne l'arrêt du fonctionnement de la partie du corps que le cerveau lésé contrôle. Un AVC est également appelé accident vasculaire cérébral (AVC) ou "attaque cérébrale".

Les types d'AVC sont les suivants :

L'accident vasculaire cérébral ischémique (une partie du cerveau perd sa circulation sanguine).

Accident vasculaire cérébral hémorragique (une hémorragie se produit dans le cerveau).

Accident ischémique transitoire, AIT ou mini-accident vasculaire cérébral (les symptômes de l'AVC disparaissent en quelques minutes, mais peuvent prendre jusqu'à 24 heures sans traitement. Il s'agit d'un signe d'alerte indiquant qu'un AVC peut survenir dans un avenir proche).

Questions à poser :

1) Hommes/femmes qui ont le plus d'accidents vasculaires cérébraux.

2) Quel groupe d'âge est le plus susceptible de subir un AVC ?

3) L'hypertension est-elle une cause ?

4) Une personne souffrant d'une maladie cardiaque est plus susceptible de subir un AVC (à confirmer).

5) Le mariage peut être une cause d'AVC.

6) Les personnes travaillant dans le secteur privé peuvent être la majorité des personnes victimes d'un AVC (principalement à cause du stress).

7) Les personnes vivant dans des zones urbaines ont plus de chances de subir un AVC (à confirmer).

8) Les niveaux de glucose sont importants et doivent être observés de près avec d'autres éléments.

9) L'IMC doit être observé de près avec l'âge et le sexe.

10) Les personnes qui fument ont plus de chances de subir un AVC (à confirmer).

Objectif :

Faire ressortir après notre analyse les principaux facteurs pour lesquels un patient est risqué d'être victime d'un AVC en fonction de paramètres d'entrée tels que le sexe, l'âge, diverses maladies et le tabagisme....

Choix de la méthode d'analyse

L'analyse factorielle des données mixtes (AFDM) est une méthode en composantes principales dédiée à l'analyse d'un ensemble de données contenant à la fois des variables quantitatives et qualitatives. Elle permet d'analyser la similarité entre les individus en prenant en compte un mélange de types de variables. De plus, on peut explorer l'association entre toutes les variables, qu'elles soient quantitatives ou qualitatives.

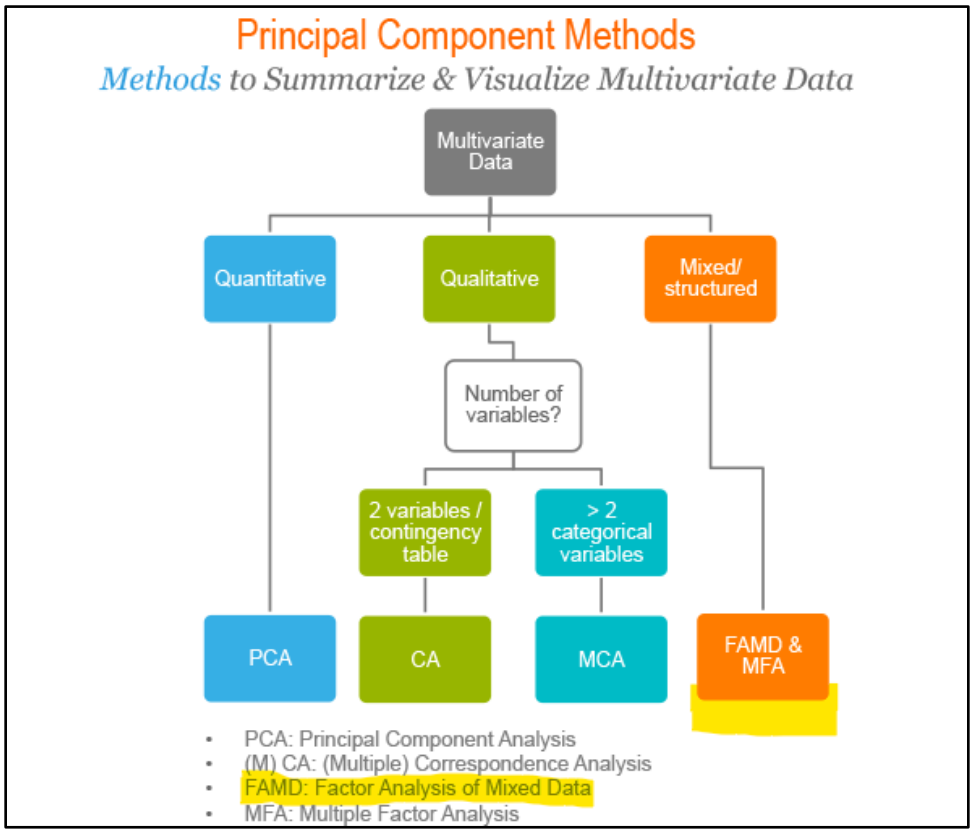


Figure 1 : Méthodes d'Analyse

En général, l'algorithme AFDM peut être considéré comme un mélange entre l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM).

En d'autres termes, il agit comme ACP pour les variables quantitatives et comme ACM pour les variables qualitatives.

Les variables quantitatives et qualitatives sont normalisées pendant l'analyse afin d'équilibrer l'influence de chaque ensemble de variables.

Informations sur le jeu de données et liste des variables

La base de données représente des observations médicales sur des patients dédié à l'étude de la présence de l'AVC.

id: Patient ID
gender : Sexe du patient
age : Age du patient
hypertension : 0 - pas d'hypertension, 1 hypertension
heart-disease : 0 - pas de maladie cardiaque, 1 pas de maladie cardiaque
ever_married : oui ou non
work_type : Type de profession
Residence_type : Sont type de résidence (Urbain / Rural)
avg_glucose_level : Niveau moyen de glucose (mesure après le repas)
bmi : indice de masse corporelle
smoking_status : statut tabagique du patient
Stoke : 0 - pas d'AVC, 1 – ayant subi un AVC

Processus d'analyse de l'AFDM:

Les librairies utilisées :

```
library('xlsx') : Pour lire les fichiers Excels
library('ade4') : Pour l'analyse des données
library('glue') : Pour l'amélioration, le formatage et l'affichage des
données
library('FactoMineR') : Pour l'analyse des données
library('factoextra') : Extraire et visualiser les résultats des analyses de données
multivariées
```

Chargement des données : "AVC_data.csv"

```
AVC_donnees <- read.table(file.choose(),fill = TRUE,
header=TRUE,sep=",",dec=".", row.names =1,stringsAsFactors = TRUE)
```

Affichage des données :

```
View(AVC_donnees)
```

Structure des données :

```
str(AVC_donnees)
```

```
'data.frame':   29072 obs. of  11 variables:
 $ gender           : Factor w/ 3 levels "Female","Male",..: 2 1 1 1 1 1 2 1 1 1 ...
 $ age              : int  58 70 52 75 32 74 79 37 37 40 ...
 $ hypertension     : int  1 0 0 0 0 1 0 0 0 0 ...
 $ heart_disease    : int  0 0 0 1 0 0 1 0 0 0 ...
 $ ever_married     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ work_type        : Factor w/ 5 levels "children","Govt_job",..: 4 4 4 5 4 5 4 4 4 4 ...
 $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 1 2 2 1 1 1 ...
 $ avg_glucose_level: num  88 69 77.6 243.5 77.7 ...
 $ bmi              : num  39.2 35.9 17.7 27 32.3 54.6 22 39.4 26.1 42.4 ...
 $ smoking_status   : Factor w/ 3 levels "formerly smoked",..: 2 1 1 2 3 2 1 2 1 2 ...
 $ stroke           : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
glue("{ncol(AVC_donnees)} variables et {nrow(AVC_donnees)} observations")
```

```
## 11 variables et 29072 observations
```

Affichage des 20 premières lignes :

```
print(AVC_donnees[1:20,])
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
30468	Male	58	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
56543	Female	70	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
52800	Female	52	0	0	Yes	Private	Urban	77.59	17.7	formerly smoked	0
41413	Female	75	0	1	Yes	Self-employed	Rural	243.53	27.0	never smoked	0
15266	Female	32	0	0	Yes	Private	Rural	77.67	32.3	smokes	0
28674	Female	74	1	0	Yes	Self-employed	Urban	205.84	54.6	never smoked	0
64908	Male	79	0	1	Yes	Private	Urban	57.08	22.0	formerly smoked	0
63884	Female	37	0	0	Yes	Private	Rural	162.96	39.4	never smoked	0
37893	Female	37	0	0	Yes	Private	Rural	73.50	26.1	formerly smoked	0
67855	Female	40	0	0	Yes	Private	Rural	95.04	42.4	never smoked	0
25774	Male	35	0	0	No	Private	Rural	85.37	33.0	never smoked	0
19584	Female	20	0	0	No	Private	Urban	84.62	19.7	smokes	0
24447	Female	42	0	0	Yes	Private	Rural	82.67	22.5	never smoked	0
49589	Female	44	0	0	Yes	Govt_job	Urban	57.33	24.6	smokes	0
17986	Female	79	0	1	Yes	Self-employed	Urban	67.84	25.2	smokes	0
72911	Female	57	1	0	Yes	Private	Rural	129.54	60.9	smokes	0
47175	Female	49	0	0	Yes	Private	Rural	60.22	31.5	smokes	0
4057	Male	71	0	0	Yes	Private	Urban	198.21	27.3	formerly smoked	0
48588	Female	59	0	0	Yes	Private	Urban	109.82	23.7	never smoked	0
70336	Female	25	0	0	Yes	Private	Urban	60.84	24.5	never smoked	0

Combien de patients ont un AVC ?

	stroke	n	proportion
1	no stroke	28524	98.115025
2	stroke	548	1.884975

La proportion des malades par âge et par sexe :

	gender	stroke	n	percentage	age_mean	age_sd
1	Female	0	17539	60.32952669	46.68727	18.55263
2	Female	1	313	1.07663731	68.71565	12.28034
3	Male	0	10978	37.76141992	48.21689	18.67442
4	Male	1	235	0.80833792	68.20000	11.25417
5	Other	0	7	0.02407815	29.28571	17.21157

Fonction pour centrage-réduction :

```
centrage_reduction <- function(x){
  n <- length(x)
  m <- mean(x)
  v <- (n-1)/n*var(x)
  return((x-m)/sqrt(v))
}
```

Appliquer la fonction sur les variables continues :

```
AVC_cr_varCont <-
data.frame(lapply(subset(AVC_donnees,select=c(2,3,4,8,9,11)),centrage_reducti
on))
summary(AVC_cr_varCont)
```

```
##       age            hypertension     heart_disease     avg_glucose_level
##  Min.   :-2.01086   Min.   :-0.3542   Min.   :-0.2346   Min.   :-1.1353
##  1st Qu.:-0.83653   1st Qu.:-0.3542   1st Qu.:-0.2346   1st Qu.:-0.6357
##  Median : 0.01752   Median :-0.3542   Median :-0.2346   Median :-0.3153
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.76482   3rd Qu.:-0.3542   3rd Qu.:-0.2346   3rd Qu.: 0.1658
##  Max.   : 1.83239   Max.   : 2.8231   Max.   : 4.2634   Max.   : 4.0790
##       bmi              stroke
```

```
##  Min.   :-2.7738   Min.   :-0.1386
##  1st Qu.:-0.7026   1st Qu.:-0.1386
##  Median :-0.1604   Median :-0.1386
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.5346   3rd Qu.:-0.1386
##  Max.   : 8.6110   Max.   : 7.2146
```

Codage disjonctif complet :

```
AVC_disjonctif <- acm.disjonctif(subset(AVC_donnees,select=c(1,5,6,7,10)))
View(AVC_disjonctif)
```

Fonction pour pondération des indicatrices :

```
fonction_ponderation <- function(x){
  m <- mean(x)
  return(x/sqrt(m))
}
```

Appliquer la pondération sur les indicatrices :

```
AVC_disjonctif_pond <-
data.frame(lapply(AVC_disjonctif,fonction_ponderation))
```

Données transformées envoyées à l'ACP :

```
AVC_pour_acp <- cbind(AVC_cr_varCont,AVC_disjonctif_pond)
rownames(AVC_pour_acp) <- rownames(AVC_donnees)

acp_AVC <- dudi.pca(AVC_pour_acp,center=T, scale=F, scannf=F)
```

Valeurs propres :

```
 [1] 2.25662 1.18962 1.10706 1.08539 1.03462 1.00580 0.99793 0.99172 0.97715 0.93695 0.86813 0.85810 0.84424 0.77686 0.69600
[16] 0.37382
```

Coordonnées ACP des variables :

```
print(acp_AVC$co[,1:2])
```

```
##                                    Comp1       Comp2
## age                          -0.8273336987  0.02986923
## hypertension                 -0.4311701084 -0.17347383
## heart_disease                -0.3926916150 -0.45163614
## avg_glucose_level            -0.4239250743 -0.19293648
## bmi                          -0.3237453135  0.28032524
## stroke                       -0.2395217153 -0.34750616
## gender.Female                 0.0777290174  0.24096294
```

```
## gender.Male                    -0.0987646834 -0.30146550
## gender.Other                    0.0275365609 -0.10309741
## ever_married.No                 0.5878987265 -0.27850846
## ever_married.Yes               -0.3429728670  0.16247840
## work_type.children              0.4094563226 -0.54450456
## work_type.Govt_job             -0.0590354371  0.18817464
## work_type.Never_worked          0.1405629393 -0.14575305
## work_type.Private               0.1173790478  0.12948138
## work_type.Self.employed        -0.3315721245 -0.20828538
## Residence_type.Rural           -0.0006496927 -0.03557686
## Residence_type.Urban            0.0006470610  0.03543275
## smoking_status.formerly.smoked -0.2563728405 -0.15056574
## smoking_status.never.smoked     0.1731344899  0.02266004
## smoking_status.smokes          -0.0015879653  0.12473809
```

Pour les valeurs qualitatives des calculs supplémentaires nécessaires ; récupérer les coordonnées ACP des modalités :

```
moda <- acp_AVC$co[7:21,1:2]
```

Fréquence des modalités :

```
freq_moda <- colMeans(AVC_disjonctif)
```

Calcul des moyennes conditionnelles sur les 2 premiers facteurs :

```
coord_moda <- moda[,1]*sqrt(acp_AVC$eig[1]/freq_moda)
coord_moda <- cbind(coord_moda,moda[,2]*sqrt(acp_AVC$eig[2]/freq_moda))
print(coord_moda)
```

```
##                                  coord_moda
## gender.Female                   0.149006939  0.33538855
## gender.Male                    -0.238895252 -0.52944168
## gender.Other                    2.665799198 -7.24669758
## ever_married.No                 1.752597762 -0.60282712
## ever_married.Yes               -0.596481175  0.20516689
## work_type.children              4.222131478 -4.07662111
## work_type.Govt_job             -0.233432756  0.54023755
## work_type.Never_worked          3.582423267 -2.69710820
## work_type.Private               0.218377001  0.17490342
## work_type.Self-employed        -1.177270120 -0.53694715
## Residence_type.Rural           -0.001383042 -0.05498823
## Residence_type.Urban            0.001371860  0.05454365
## smoking_status.formerly smoked -0.779363712 -0.33232956
## smoking_status.never smoked     0.353387702  0.03358173
## smoking_status.smokes          -0.005154698  0.29399229
```

Carré des corrélations du 1er facteur :

```
r2 <- acp_AVC$co[1:6,1]^2
```

Carré du rapport de corrélation (Variables qualitatives) :

```
eta2 <- NULL
eta2[1] <- sum(acp_AVC$co[7:9,1]^2)
eta2[2] <- sum(acp_AVC$co[10:11,1]^2)
eta2[3] <- sum(acp_AVC$co[12:16,1]^2)
eta2[4] <- sum(acp_AVC$co[17:18,1]^2)
eta2[5] <- sum(acp_AVC$co[19:21,1]^2)
print(eta2)
```

```
## [1] 1.655453e-02 4.632553e-01 3.146155e-01 8.407886e-07 9.570511e-02
```

Critère de l'AFDM au 1er facteur :

```
lambda1 <- sum(criteres)
print(lambda1)
```

```
## [1] 2.256621
```

Confrontation avec résultat (v.p) de l'ACP sur variables transformées au 1er facteur :

```
print(acp_AVC$eig[1])
```

```
## [1] 2.256621
```

AFDM avec FactoMineR

Lancement de la procédure sur 10000 individus :

```
afdm_AVC_donnee <- FAMD(AVC_donnees[1:10000,], ncp = 2)
```

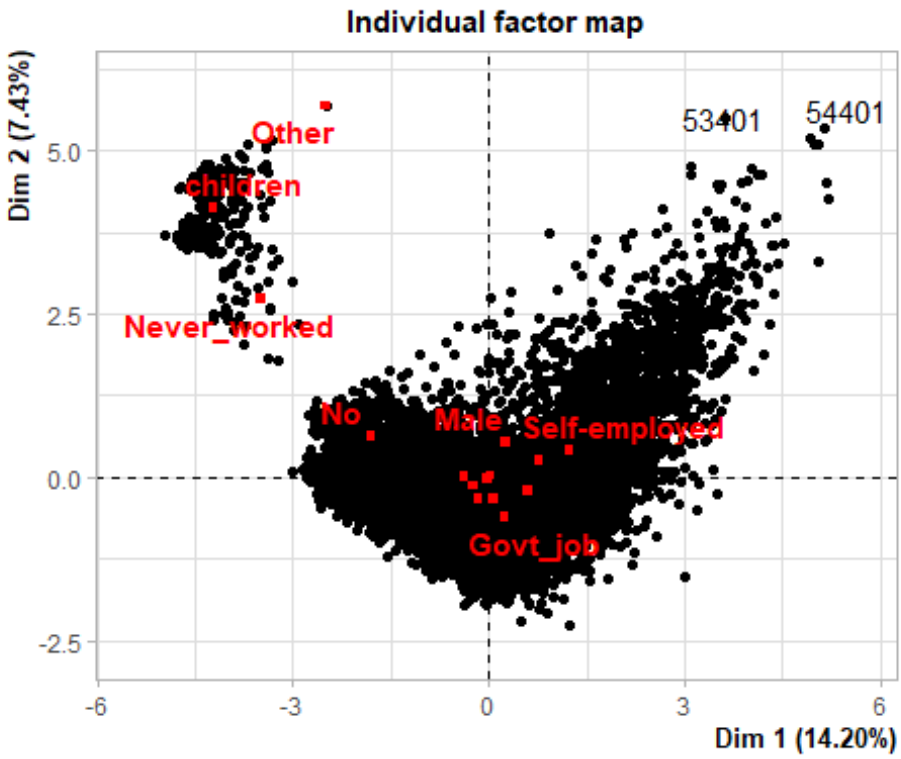


Figure 2 : Graphe des individus

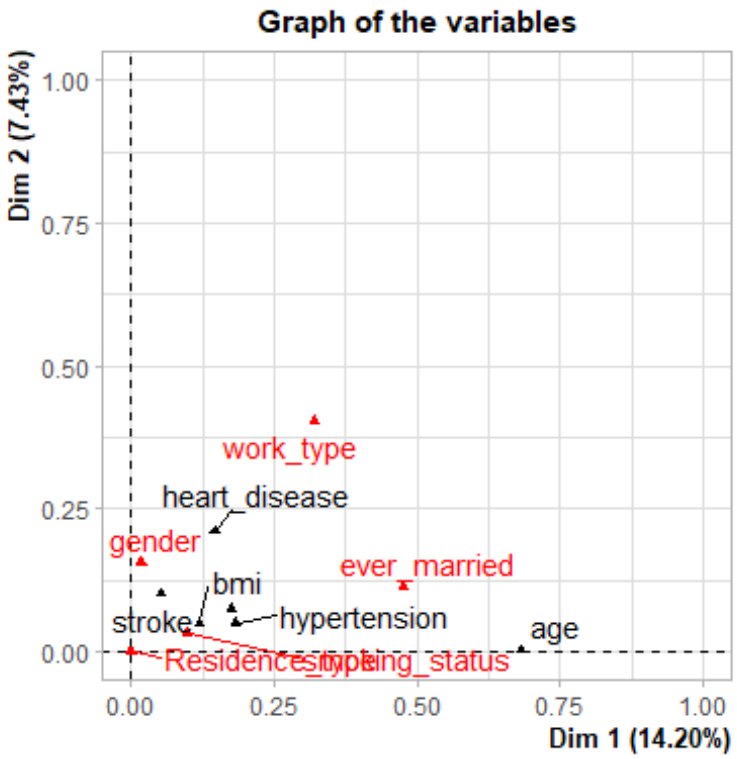


Figure 3 : Graphe des variables

Le graphe ci-dessus représente le carré des corrélations des variables quantitatives (noir) ou carré des rapports de corrélation des qualitatives (rouge). Il montre que la variable « age » est parmi les plus déterminants sur la première dimension et «work_type » est la variable qui pèse le plus sur la deuxième dimension.

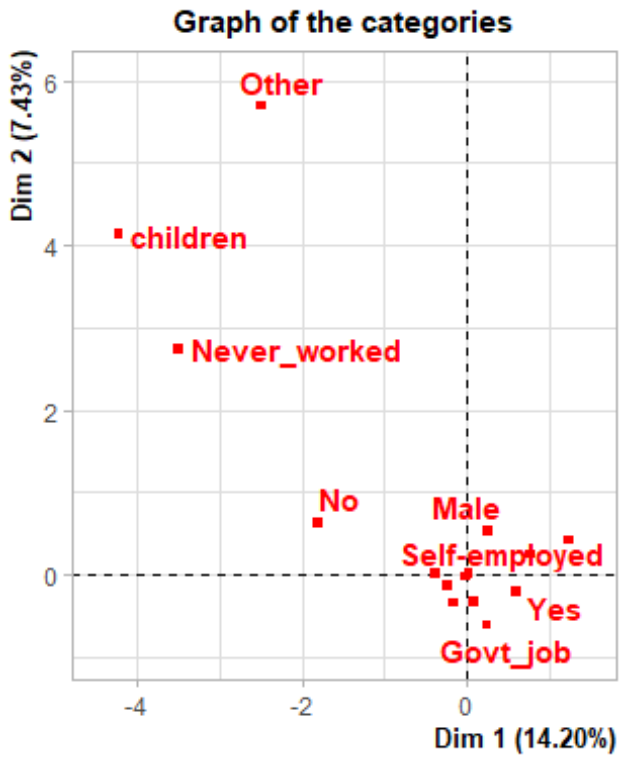


Figure 4 : Graphe des modalités

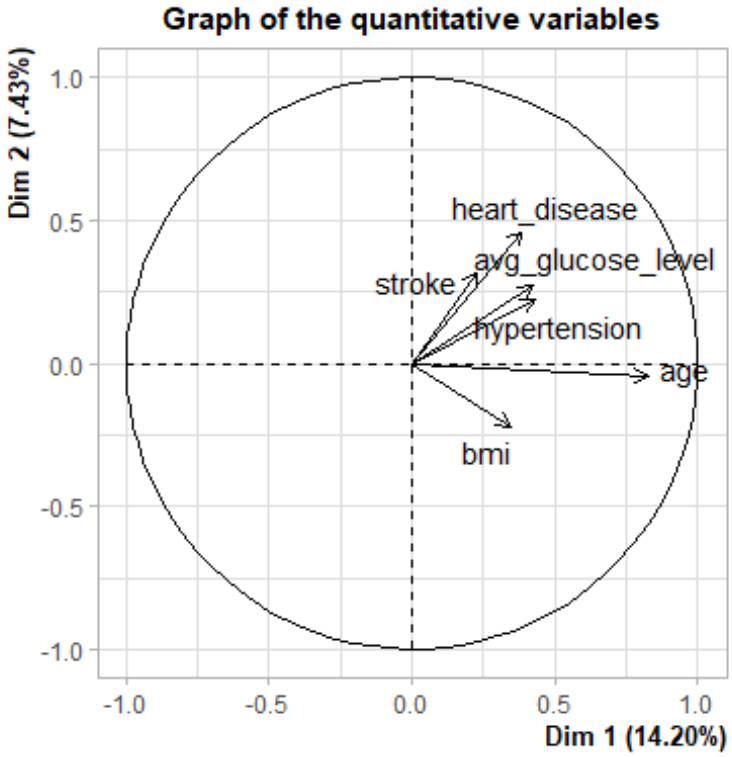


Figure 5 : Cercle de corrélation des variables quantitatives

Les variables **hear_disease, avg_glucose_level, hypertension** et **stroke** sont dans la même direction donc ils sont corrélés cela signifie que à la présence de maladie cardiaque et/ou de l'hypertension et/ou un niveau élevé de la moyenne de glucose le risque d'AVC augmente.

La variable **stroke** est perpendiculaire à l'indice de la masse corporelle **bmi**. On peut déduire qu'il n'y a pas de lien entre le **bmi** et **l'AVC.**

La variable "age" est trés proche de l'axe de la première dimension cela veut dire qu'il est très bien expliqué par cet axe.

Lancement de la procédure sur tous les individus (analyse général) :

```
afdm_AVC_donnees <- FAMD(AVC_donnees, graph = FALSE)
```

Visualisation des données Traçage des variables :

```
viz1 <- fviz_famd_var(afdm_AVC_donnees, repel = TRUE)
viz1
```

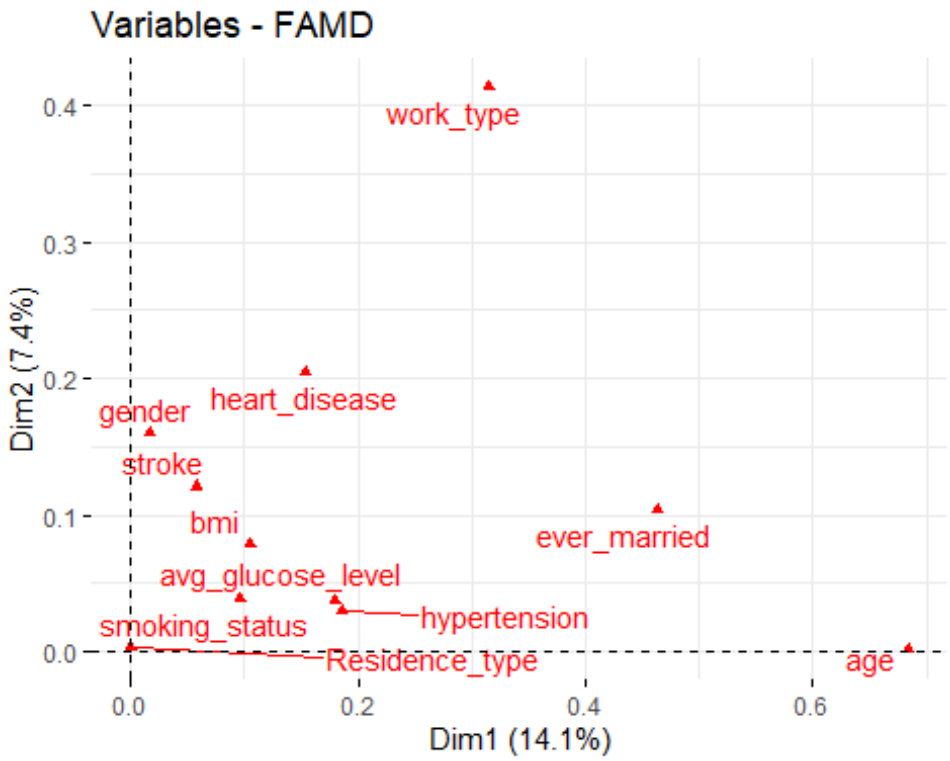


Figure 6 : variable AFMD

Contribution de la première dimension :

```
viz2 <- fviz_contrib(afdm_AVC_donnees, "var", axes = 1)
viz2
```

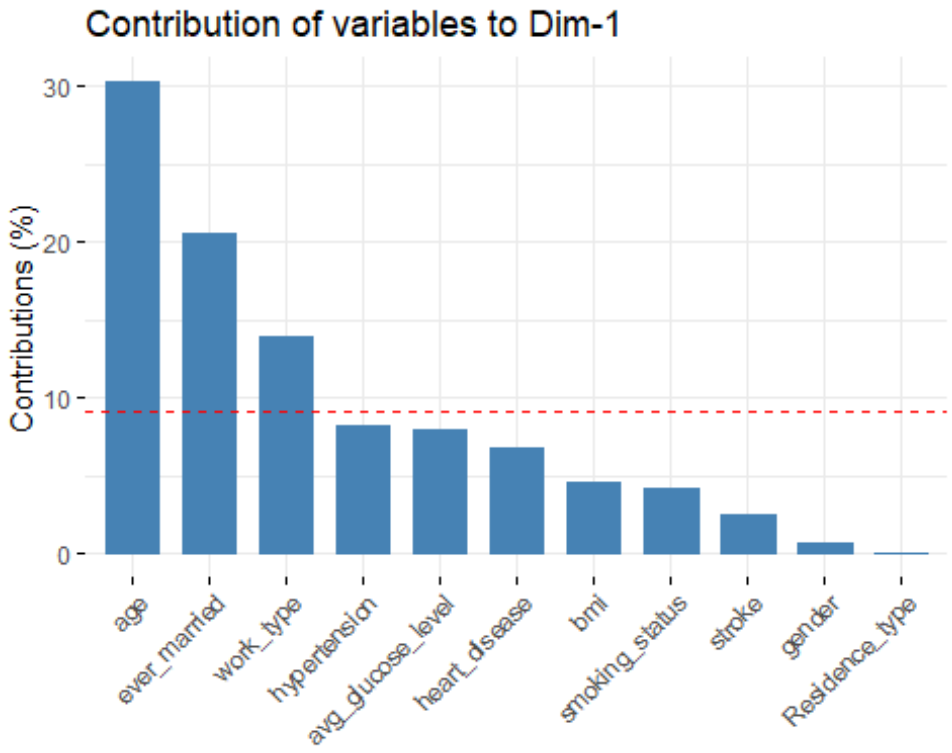


Figure 7 : Contribution des variables à la première Dimension

```
viz3 <- fviz_contrib(afdm_AVC_donnees, "var", axes = 2)
viz3
```

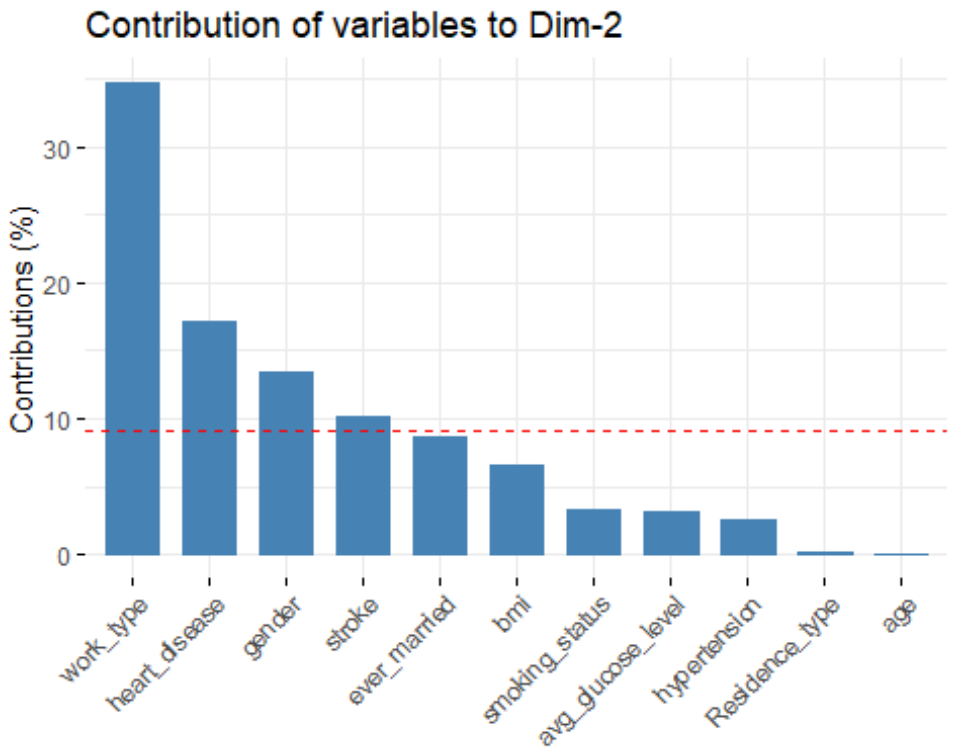


Figure 8 : Contribution des variables à la Deuxième Dimension

De cette réduction de dimension, il ressort que l'âge, l'état matrimonial et le type de travail contribuent le plus à la première dimension, tant dit que le type de travail, les maladies cardiaques, le sexe et les AVC contribuent le plus à la deuxième dimension.

Cette conclusion confirme les interprétations visuelles de la figure 3.

```
quali.var <- get_famd_var(afdm_AVC_donnees, "quali.var")
quanti.var <- get_famd_var(afdm_AVC_donnees, "quanti.var")
fviz_famd_var(afdm_AVC_donnees, "quanti.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE)
```

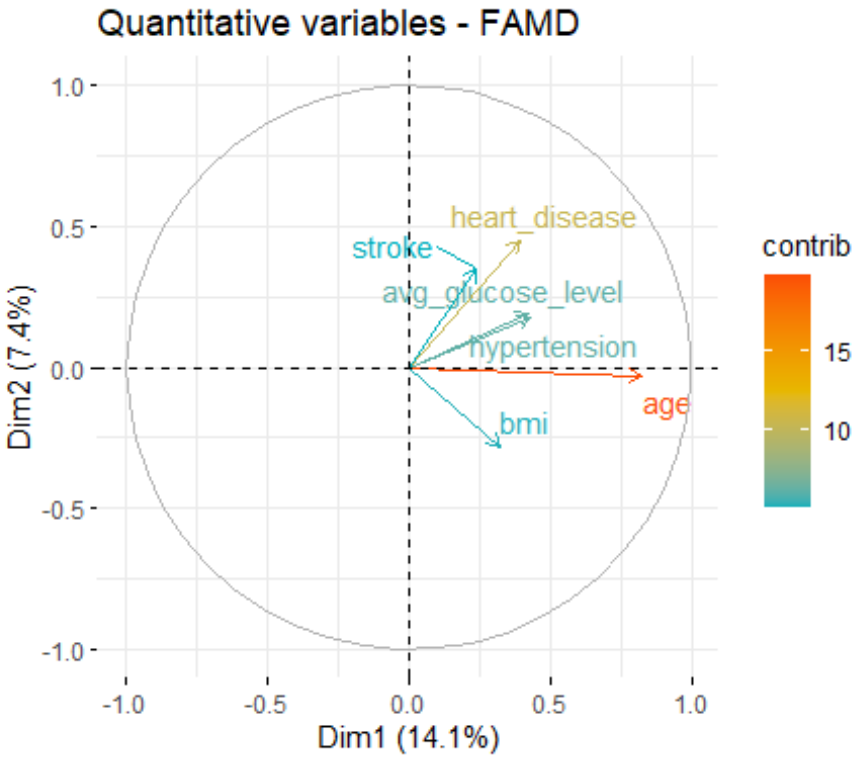


Figure 9 : Cercle de corrélation des variables quantitatives avec contribution

```
fviz_famd_var(afdm_AVC_donnee, "quali.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE
)
```

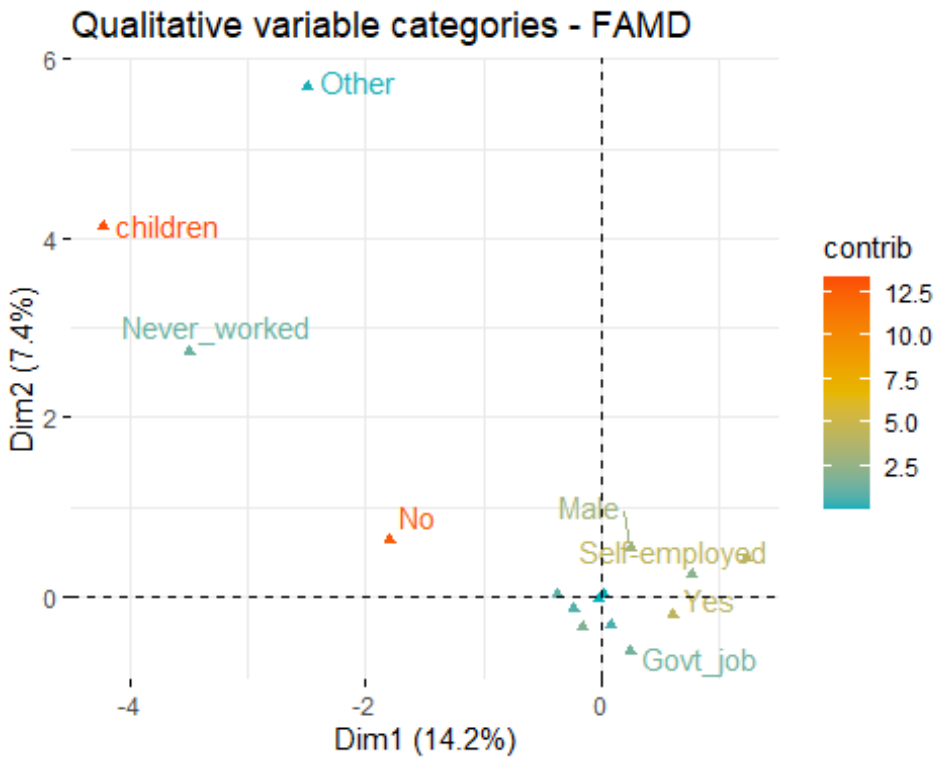


Figure 10 : Modalités des variables qualitatives avec contribution

Le graphique ci-dessus montre la corrélation des modalités des variables qualitatives sur l'espace représenté par les 2 dimensions ainsi que leurs contributions sur ces 2 axes.

```
famd.stroke <- fviz_mfa_ind(afdm_AVC_donnees,
                            habillage = "stroke", # color by groups
                            geom = c('point'),
                            palette = pal_simpsons("springfield", alpha =
0.6)(16),
                            title = "Stroke Status"
)
famd.stroke
```

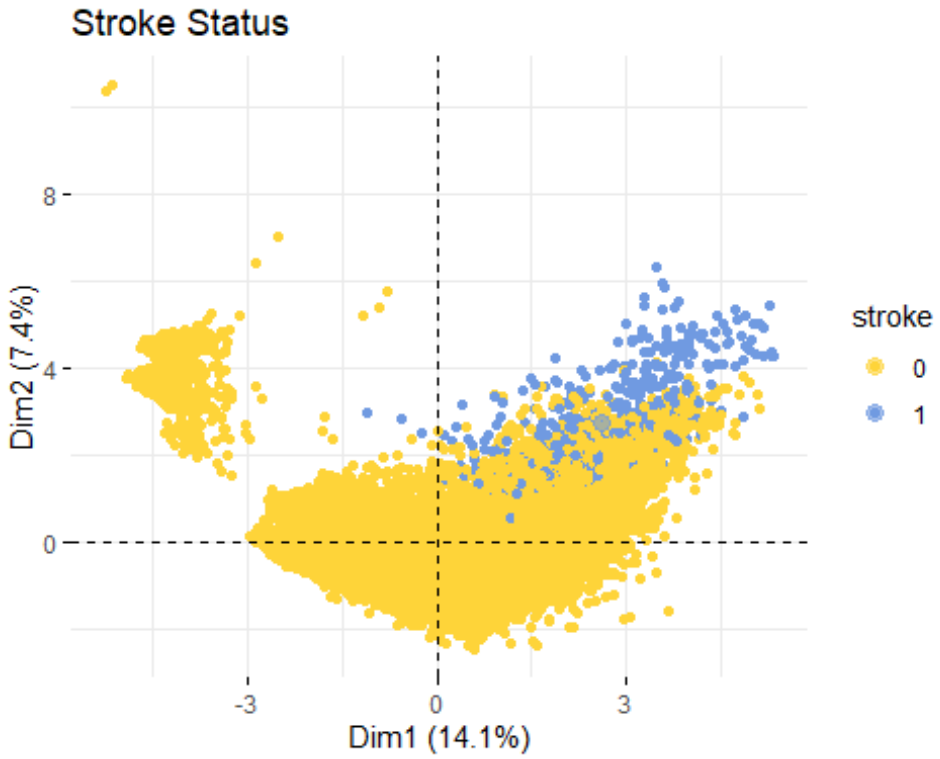


Figure 11 : Status d'AVC

0 => Patient n'ayant pas d'AVC, 1 => Patient ayant subit un AVC

La figure ci-dessus confirme visuellement les interprétations faites sur la figure 8.

```
famd.smoking <- fviz_mfa_ind(afdm_AVC_donnees,
                             habillage = "smoking_status", # color by groups
                             geom = c('point'),
                             palette = pal_simpsons("springfield", alpha =
0.6)(16),
                             title = "Smoking Status"
)
famd.smoking
```

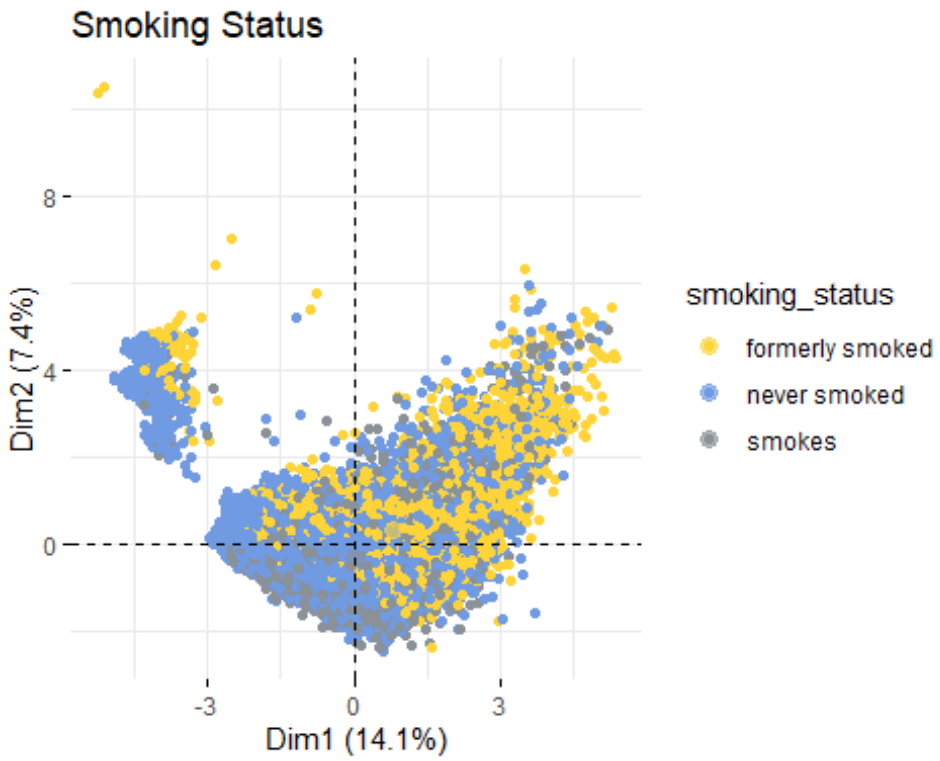


Figure 12 : Statut tabagique

La figure ci-dessus confirme que les modalités de la variable “smoking_status” n’aident pas à la description de l’une des axes.

```
famd.married <- fviz_mfa_ind(afdm_AVC_donnees,
                             habillage = "ever_married", # color by groups
                             geom = c('point'),
                             palette = pal_simpsons("springfield", alpha =
0.6)(16),
                             title = "Martial Status"
)
famd.married
```

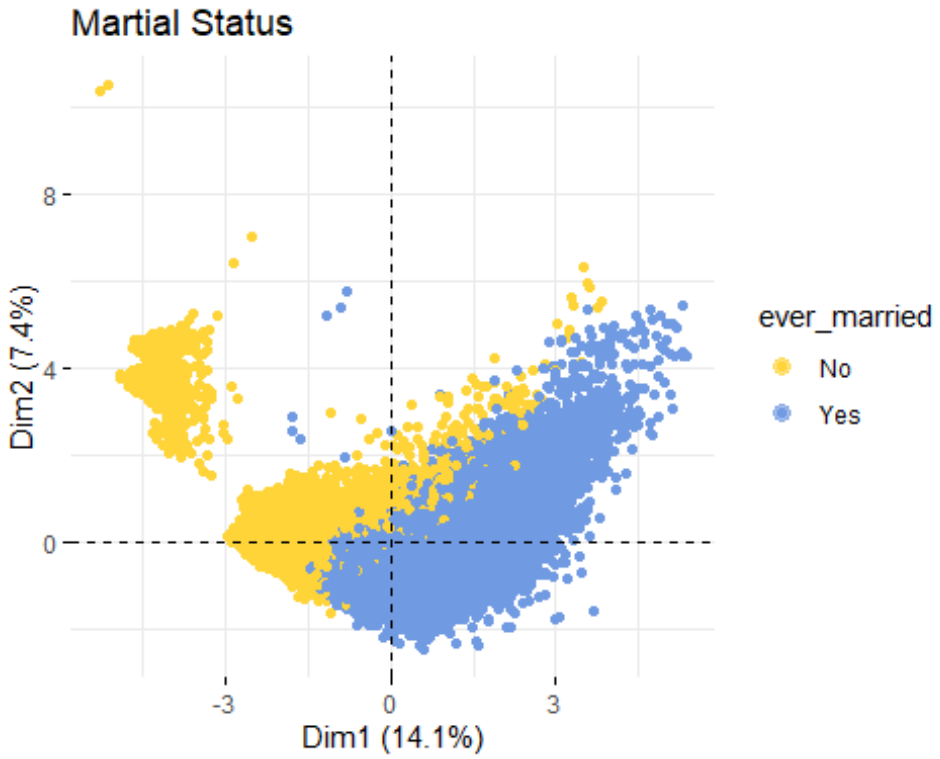


Figure 13 : Statut d'état matrimonial

La figure ci-dessus confirme visuellement les interprétations faites sur la figure 7.

```
famd.work_type <- fviz_mfa_ind(afdm_AVC_donnee,
                               habillage = "work_type", # color by groups
                               geom = c('point'),
                               palette = pal_simpsons("springfield", alpha =
0.6)(16),
                               title = "Work Status"
)
famd.work_type
```

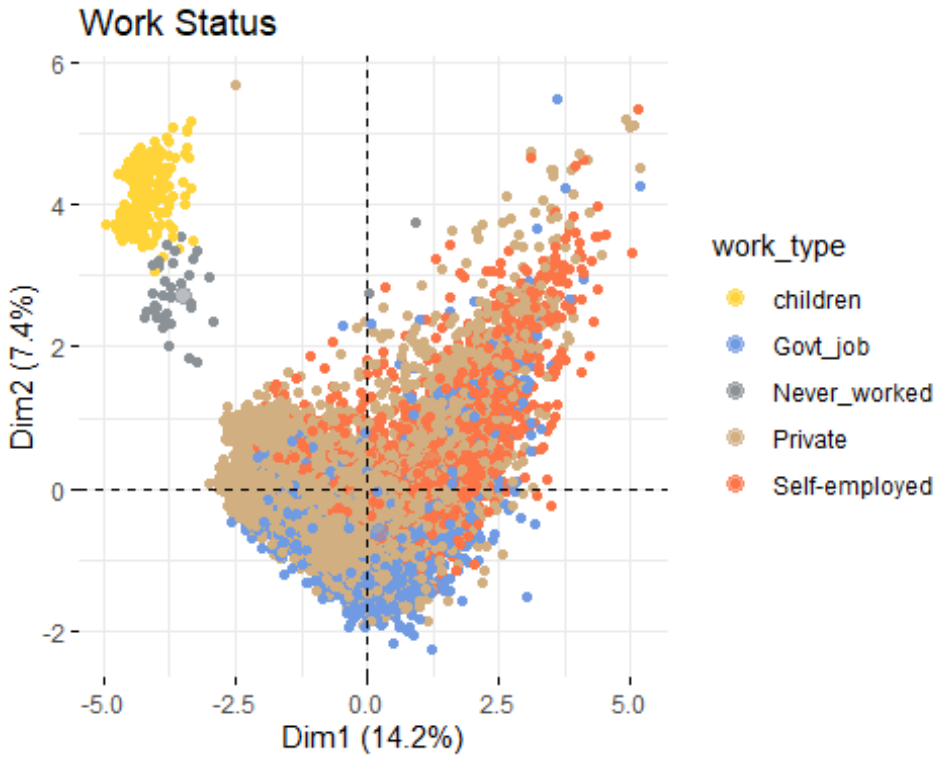
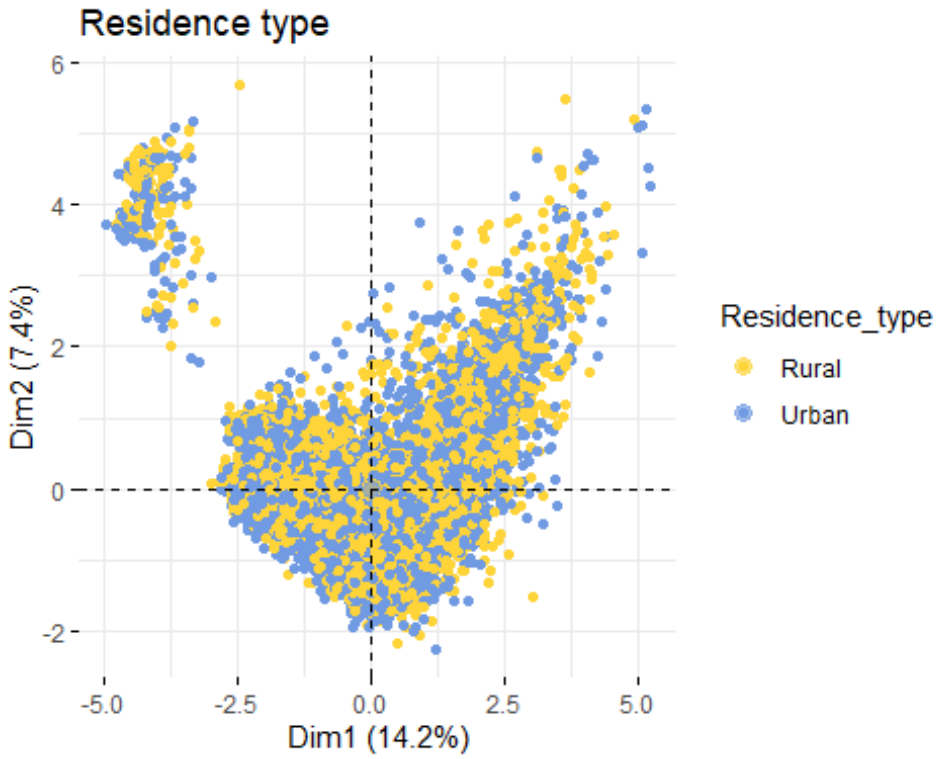


Figure 14 : Type de profession

La figure ci-dessus confirme visuellement les interprétations faites sur la figure 8.

```
famd.residence <- fviz_mfa_ind(afdm_AVC_donnee,
                               habillage = "Residence_type", # color by
groups
                               geom = c('point'), # rempove labels
                               palette = pal_simpsons("springfield", alpha =
0.6)(16), # use a color blind friendly palette
                               title = "Residence type"
)
famd.residence
```



La figure ci-dessus confirme que les modalités de la variable “Residesnce_type” n’aident pas à la description de l’un des axes.

```
famd.heart_disease <- fviz_mfa_ind(afdm_AVC_donnee,
                                   habillage = "heart_disease", # color by
groups
                                   geom = c('point'),
                                   palette = pal_simpsons("springfield",
alpha = 0.6)(16),
                                   title = "Heart Disease Status"
)
famd.heart_disease
```

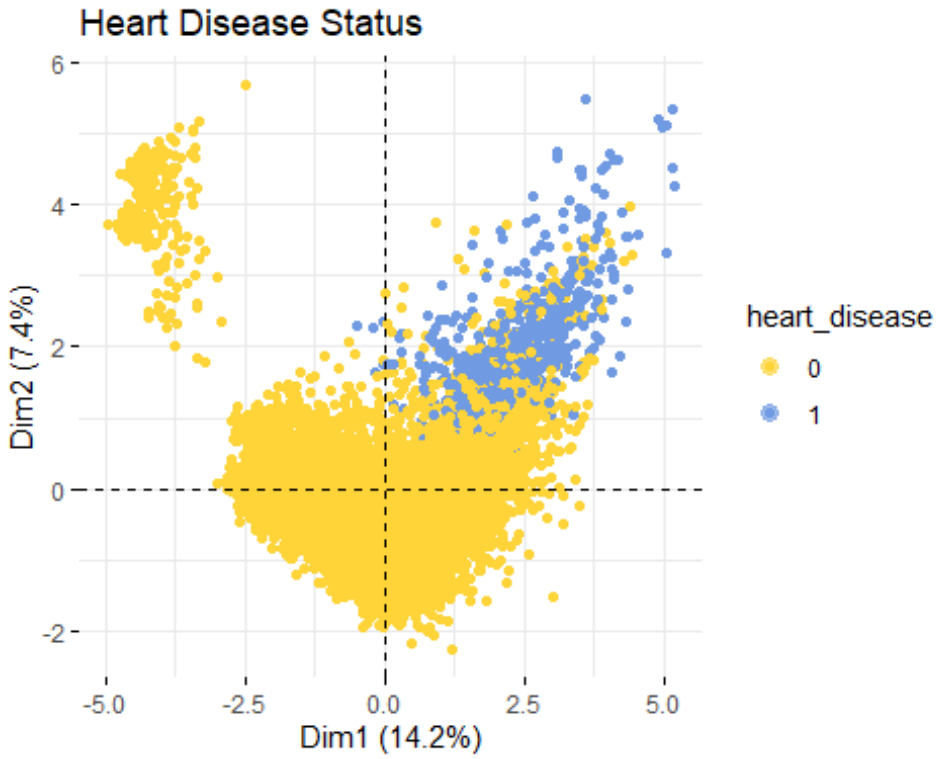


Figure 15 : Statut des maladies cardiaques

La figure ci-dessus confirme visuellement les interprétations faites sur la figure 8.

```
famd.hypertension <- fviz_mfa_ind(afdm_AVC_donnee,
                                  habillage = "hypertension", # color by
groups
                                  geom = c('point'),
                                  palette = pal_simpsons("springfield", alpha
= 0.6)(16),
                                  title = "Hypertension Status"
)
famd.hypertension
```

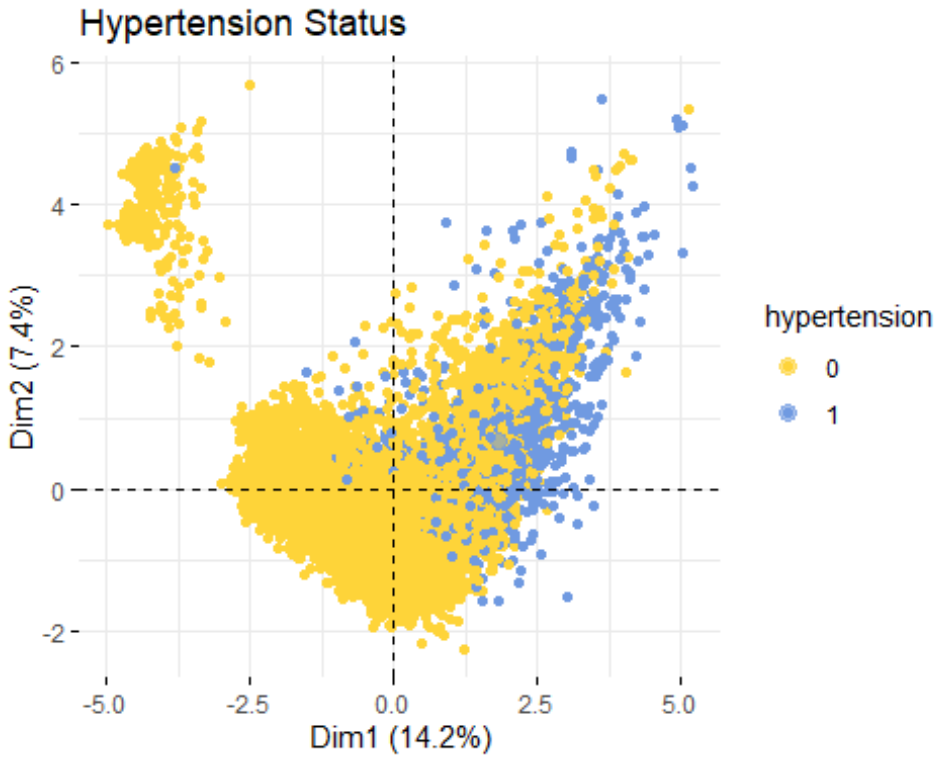
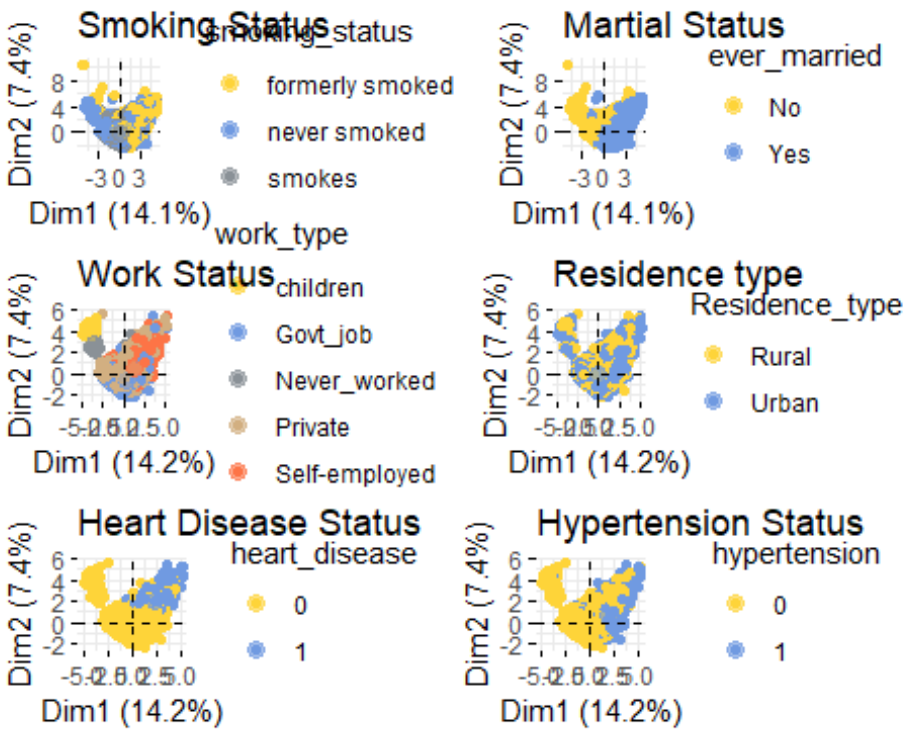


Figure 16 : Status de l'hypertension

La figure ci-dessus confirme que les modalités de la variable “hypertension” n’aident pas forcément à la description de l’un des axes.

```
all_FAMD <- (famd.smoking + famd.married) / (famd.work_type + famd.residence)
/ (famd.heart_disease + famd.hypertension)
plot(all_FAMD)
```



Conclusion

L'AFDM est une méthode de réduction de dimension qui sert à synthétiser l'analyse des données avec des variables quantitatives et qualitatives. Ces variables sont prises en compte de façon équivalente pour déterminer les dimensions de la variabilité. Cette méthode permet d'étudier les ressemblances entre individus en prenant en compte des variables mixtes et d'étudier les relations entre toutes les variables.

En utilisant cette méthode à l'aide des packages FactoMinerR et factoextra et quelques analyses supplémentaires, on a pu identifier 5 variables comme principaux facteurs de risque de développer un AVC :

- age.

- type de profession.

- L'état matrimonial.

- sexe.

- les maladies cardiaques.

Les difficultés rencontrées

Les difficultés que nous avons rencontrées à l'issue de ce travail :

*Le choix d'une base de données pertinente.

*le choix des techniques de visualisation des données.

Bibliographie

https://www.kaggle.com/code/wguesdon/stroke-prevention-clustering-and-risks-factors/data?select=train_2v.csv

http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/76-afdm-analyse-factorielle-des-donnees-mixtes-avec-r-l-essentiel/

https://www.kaggle.com/code/wguesdon/stroke-prevention-clustering-and-risks-factors/script

https://www.eyrolles.com/Informatique/Livre/apprentissage-statistique-9782212122299/#:~:text=L'apprentissage%20statistique%20permet%20la,en%20environnement%20complexe%20et%20%C3%A9volutif.

https://www.datanovia.com/en/product/practical-guide-to-principal-component-methods-in-r/