IA SCHOOL

Rapport du projet

Data Appliqué à la Santé

Marie-Ange Dieng
25/03/2025

Table des matières

INTRODUCTION 3
A. Contexte 3
B. Problématique 3
C. Jeu de donnée 3
I. Préparation du jeu de donnée 3
1. Importation et exploration 3
2. Preparation des variables 4
II. Analyse Descriptive Univariee 5
A. Variables Quantitatives 5
1. Age 5
2. Taille 6
3. Poids 7
B. Variables Qualitatives 8
1. Infarctus (Variable Cible) 8
2. Tabac 9
3. Contraceptif Oral 9
4. Antécédent Familiaux 10
5. Hypertension Artérielle 11
III. Analyse Bivariée 12
A. Variables quantitatives VS Variable Infarctus 12
1. Age vs Infarctus 12
2. Taille vs Infarctus 13
3. Poids vs Infarctus 14
4. IMC vs Infarctus 14
5. Matrice de Corrélation 16
B. Variables Qualitatives vs Infarctus 16
1. Contraceptif Oral vs Infarctus 16
2. Tabac vs Infarctus 17
3. Antécédent familiaux vs Infarctus 18
4. Hypertension vs Infarctus 20
IV. Analyse Multivariée 21
A. ACP sur des données quantitatives 21
B. MCA sur variables qualitatives 22
V. Modelisation 24
A. Préparation à la modèlisation 24
B. Régression Logistique 24
C. Random Forest 27

D. SVM 30
E. Comparaison des modèles 31
Conclusion 32

INTRODUCTION

A. Contexte

Les maladies cardiovasculaires, et en particulier l'infarctus, sont une cause majeure de morbidité et de mortalité chez les femmes. Cette étude repose sur une enquête cas-témoins menée auprès de 149 femmes ayant subi un infarctus du myocarde (cas) et 300 femmes n'en ayant pas eu (témoins). L'objectif est d'évaluer l'existence d'un risque accru d'infarctus chez les femmes utilisant ou ayant utilisé des contraceptifs oraux. Outre ce facteur principal, plusieurs autres variables ont été collectées, notamment l'âge, le poids, la taille, la consommation de tabac, l'hypertension artérielle et les antécédents familiaux de maladies cardiovasculaires.

B. Problématique

L'utilisation des contraceptifs oraux pourrait-elle augmenter le risque d'infarctus du myocarde chez les femmes ? Plus largement, quels sont les facteurs explicatifs de la survenue d'un infarctus et leur importance relative ? Cette étude vise à identifier les variables influençant significativement la probabilité de développer un infarctus afin d'éclairer les décisions médicales et préventives.

C. Jeu de donnée

L'étude repose sur un ensemble de données comprenant les variables suivantes :

- **INFARCT** (*Variable cible*) : 0 = non, 1 = oui
- **CO** *(contraceptif oral)* : 0 = non, 1 = oui
- **TABAC** *(consommation de tabac)* : 0 = non, 1 = actuelle, 2 = ancienne
- **AGE** (*Âge en années*)
- **POIDS** (*Poids en kilogrammes*)
- **TAILLE** (*Taille en centimètres*)
- **ATCD** *(antécédents familiaux de maladies cardiovasculaires)* : 0 = non, 1 = oui
- **HTA** *(hypertension artérielle)* : 0 = non, 1 = oui

Cet ensemble de données sera analysé afin de détecter d'éventuelles relations significatives entre ces facteurs et la survenue d'un infarctus.

I. Préparation du jeu de donnée

1. Importation et exploration

```r
df <- read.csv("data_infarctus.csv", sep = ";", header=TRUE)

head(df)
```

```
##   NUMERO INFARCT CO TABAC AGE POIDS TAILLE ATCD HTA
## 1      1       0  0     0  47    48    173    0   0
## 2      2       0  0     0  17    NA    162    0   0
## 3      3       0  0     0  35    53    163    0   0
## 4      4       0  0     0  82    78    157    0   0
## 5      5       0  0     0  50    52    172   NA   0
## 6      6       0  0     0  31    47    184    0   0
```

-Suppression de la colonne des index « NUMERO »

```
df <- df[, !(names(df) %in% c("NUMERO"))]
```

-Verification de la liste et du type de colonne

```
str(df)
```

```
## 'data.frame':    449 obs. of  8 variables:
##  $ INFARCT: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CO     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ TABAC  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AGE    : int  47 17 35 82 50 31 60 30 44 38 ...
##  $ POIDS  : int  48 NA 53 78 52 47 60 75 68 NA ...
##  $ TAILLE : int  173 162 163 157 172 184 169 174 164 167 ...
##  $ ATCD   : int  0 0 0 0 NA 0 0 0 0 0 ...
##  $ HTA    : int  0 0 0 0 0 0 0 0 0 0 ...
```

-Statistiques initiales par colonne

```
summary(df)
```

```
##     INFARCT             CO             TABAC             AGE       
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 15.00  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 33.00  
##  Median :0.0000   Median :0.0000   Median :1.0000   Median : 44.00  
##  Mean   :0.3318   Mean   :0.4454   Mean   :0.7416   Mean   : 45.62  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 56.00  
##  Max.   :1.0000   Max.   :1.0000   Max.   :2.0000   Max.   :100.00  
##                                                                     
##      POIDS            TAILLE           ATCD             HTA        
##  Min.   : 33.00   Min.   :138.0   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.: 51.00   1st Qu.:160.0   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median : 64.00   Median :166.0   Median :0.0000   Median :0.0000  
##  Mean   : 66.07   Mean   :165.2   Mean   :0.1199   Mean   :0.3541  
##  3rd Qu.: 79.00   3rd Qu.:171.0   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :128.00   Max.   :184.0   Max.   :1.0000   Max.   :1.0000  
##  NA's   :12                       NA's   :7
```

2. Preparation des variables

-Récuperation des variables quantitatives

```
age <-df$AGE
taille <-df$TAILLE
poids <-df$POIDS
```

-Récupération des variables qualitatives et mise en facteurs

```
infarctus <-factor(df$INFARCT, levels=c(0,1), labels=c("non", "oui"))
contraceptif <- factor(df$CO, levels=c(0,1), labels=c("non", "oui"))
antecedent <- factor(df$ATCD, levels=c(0,1), labels=c("non", "oui"))
hypertension <- factor(df$HTA, levels=c(0,1), labels=c("non", "oui"))
```

```
tabac <- factor(df$TABAC, levels=c(0,1,2), labels=c("nonfumeuse", "actuelle", "ancienn
e"))
```

II. Analyse Descriptive Univariee

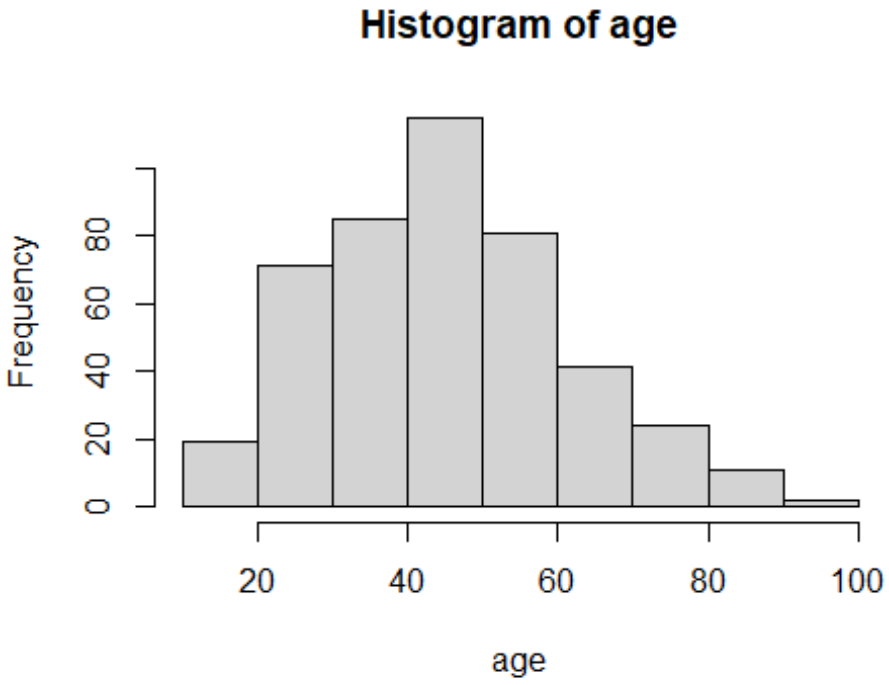
A. Variables Quantitatives

1. Age

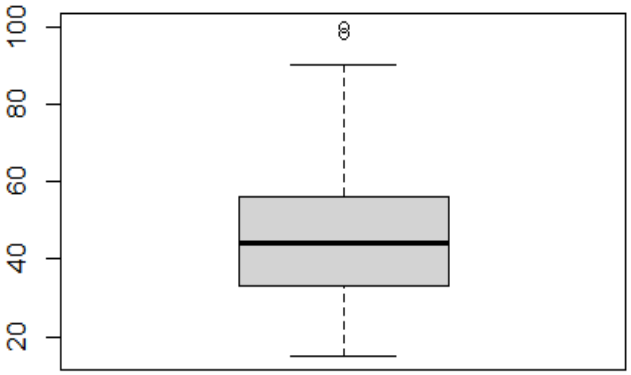
Moyenne: 45.619

Ecart-Type: 16.166

Histogramme:



BoxPlot:



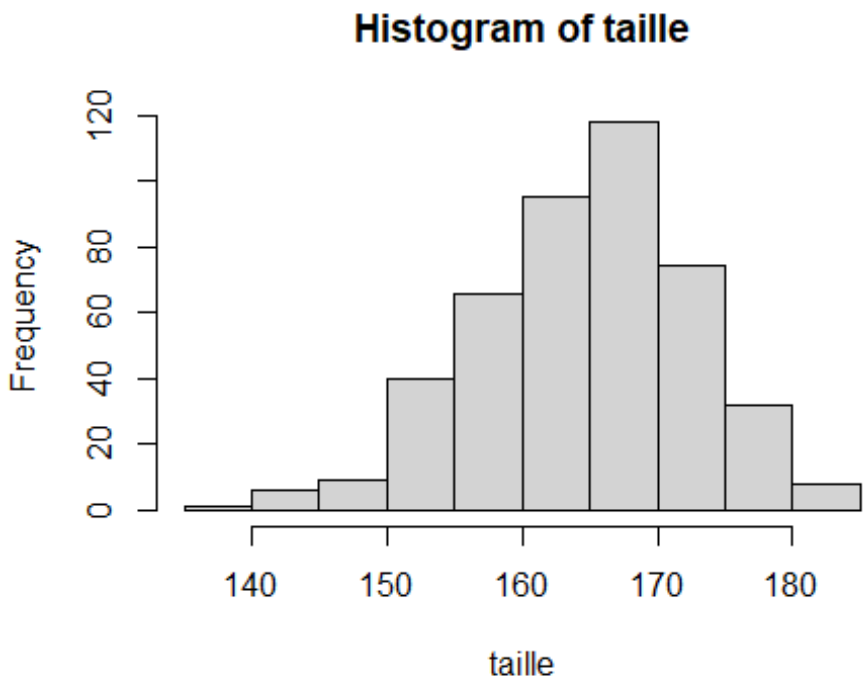
La variable âge à une médiane d'environ **45 ans**. Le premier quartile est à 30 ans, donc environ 25% des observations ont au plus **30 ans.** Le troisième quartile est à 60 ans, donc environ 75% des observations ont au plus **60 ans.** De manière générale l'âge de l'échantillon varie de **20 ans** à **85 ans**. On remarque aussi des points de valeurs aberrantes. Ici, il y a **quelques individus ayant un âge supérieur à 85 ans, jusqu'à environ 100 ans**.

2. Taille

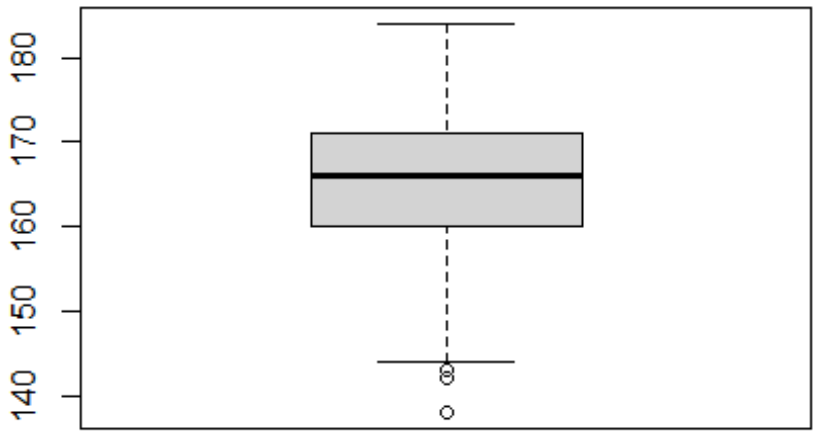
Moyenne: 165.16

Ecart-Type: 8.106

Histogramme:



BoxPlot:



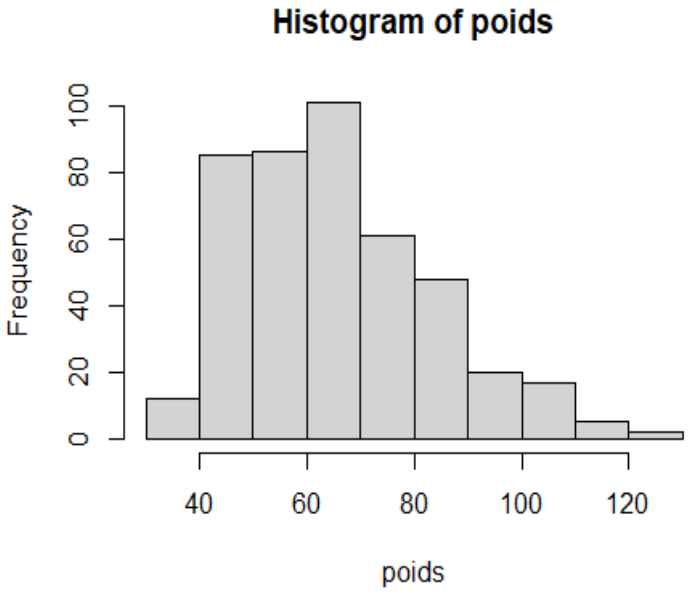
La variable taille a une médiane d'environ **165 cm**. Premier quartile (Q1) à environ **160 cm**, donc 25% des observations ont une taille inférieure ou égale à 160 cm. Troisième quartile (Q3) à environ 175 cm, donc 75% des observations ont une taille inférieure ou égale à **175 cm**. De manière générale, la taille de l'échantillon varie entre environ **145 cm** et **185 cm**. On remarque aussi la présence de valeurs aberrantes du côté des petites tailles (inférieures à **150 cm**), ce qui signifie qu'il y a quelques individus nettement plus petits que la majorité du groupe.

3. Poids

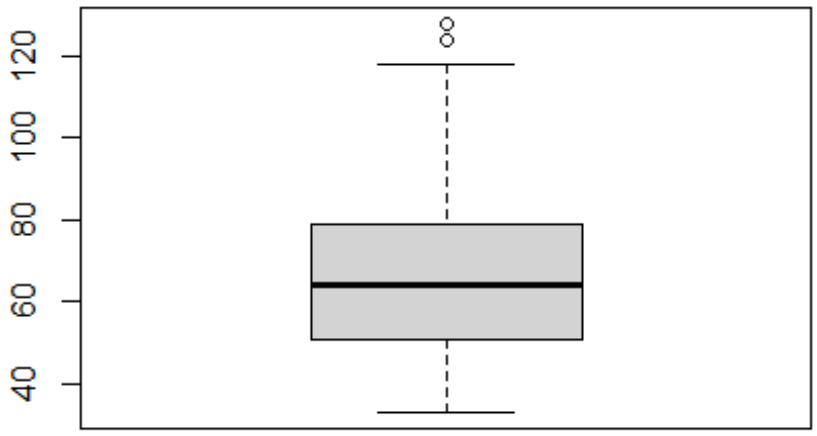
Moyenne: 66.066

Ecart-Type: 17.962

Histogramme:



BoxPlot:



La médiane se situe autour de **70 Kg**. Le premier quartile (Q1) est environ à 55 Kg, ce qui signifie que 25% des observations ont une valeur inférieure ou égale à **55 Kg**. Le troisième quartile (Q3) est environ à 85 Kg, ce qui signifie que 75% des observations ont une valeur inférieure ou égale à **85 Kg**. La majorité des valeurs sont comprises entre environ 40 et 110. On observe des valeurs aberrantes situées au-delà de **110 Kg**, atteignant environ **120-125 Kg**.

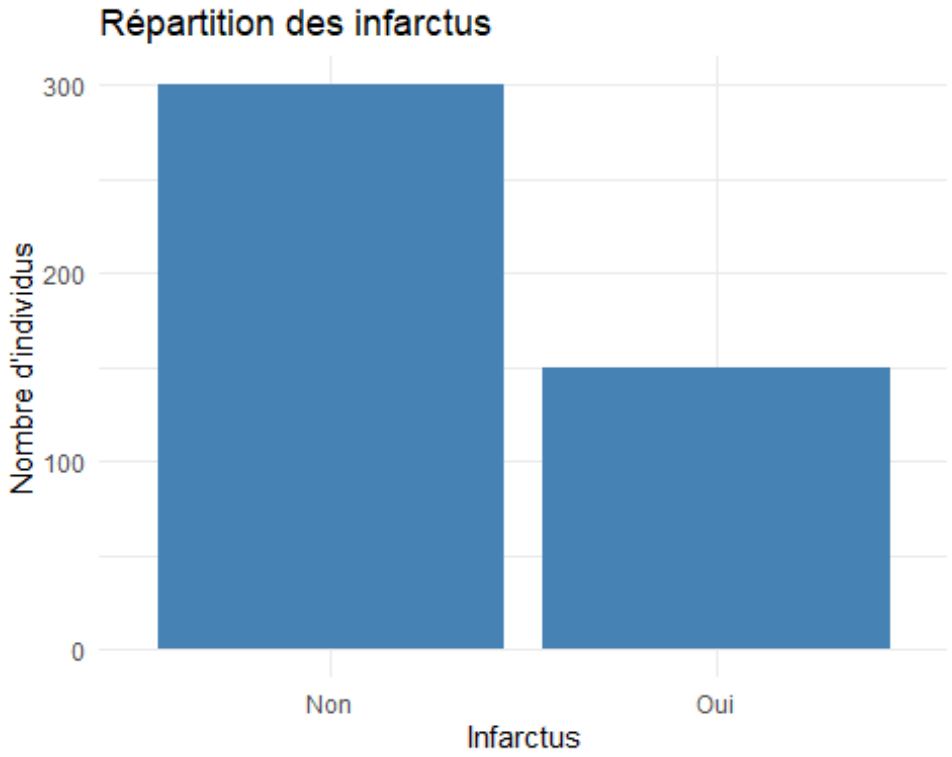
B. Variables Qualitatives

1. Infarctus (Variable Cible)

```
table(infarctus)

## infarctus
## non oui
## 300 149
```

Il y'a 300 observations de femmes n'ayant pas eu d'infarctus et 149 observations de femmes ayant eu des infarctus.

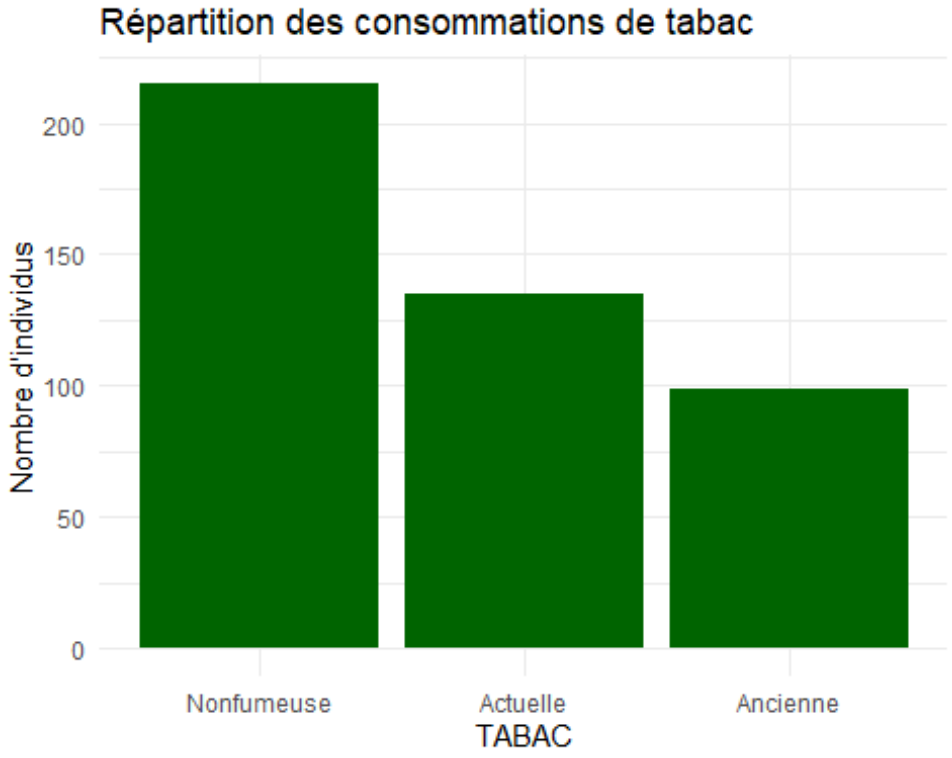


2. Tabac

```
table(tabac)

## tabac
## nonfumeuse   actuelle   ancienne
##        215        135         99
```

Parmi les obversations, on a 215 non fumeuses, 135 activement fumeuses et 99 anciennes fumeuses.

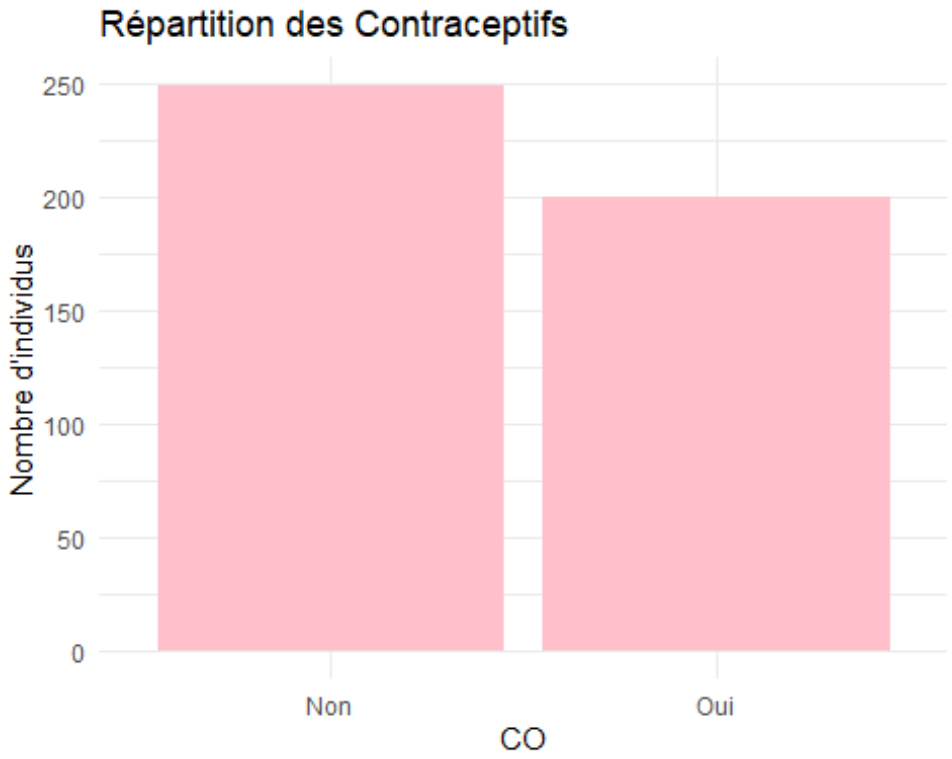


3. Contraceptif Oral

```
table(contraceptif)
```

```
## contraceptif
## non oui
## 249 200
```

Parmi les observations, on a 200 utilisatrices de contraceptifs oraux et 249 non utilisatrices.

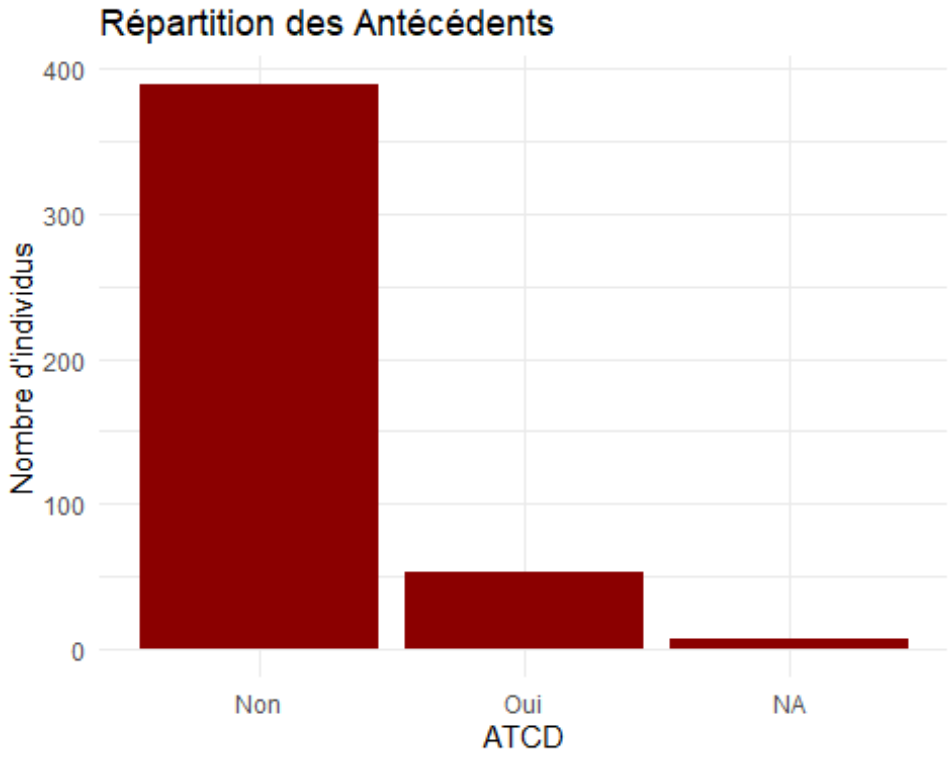


4. Antécédent Familiaux

```
table(antecedent)

## antecedent
## non oui
## 389  53
```

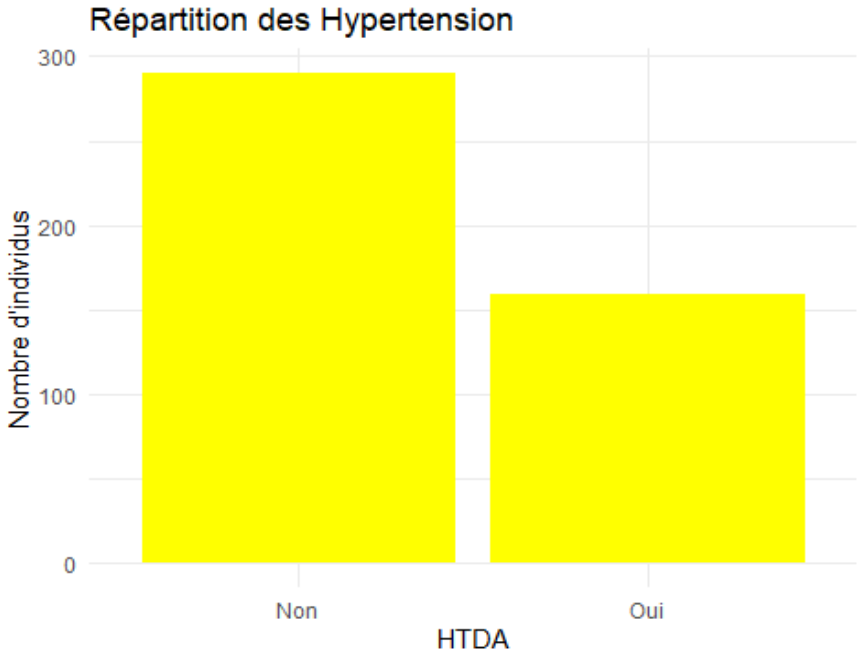
Parmi les observations, on a 53 femmes avec des antécédents familiaux de maladie cardiaque et 389 sans antécédents.



5. Hypertension Artérielle

```
table(hypertension)
```

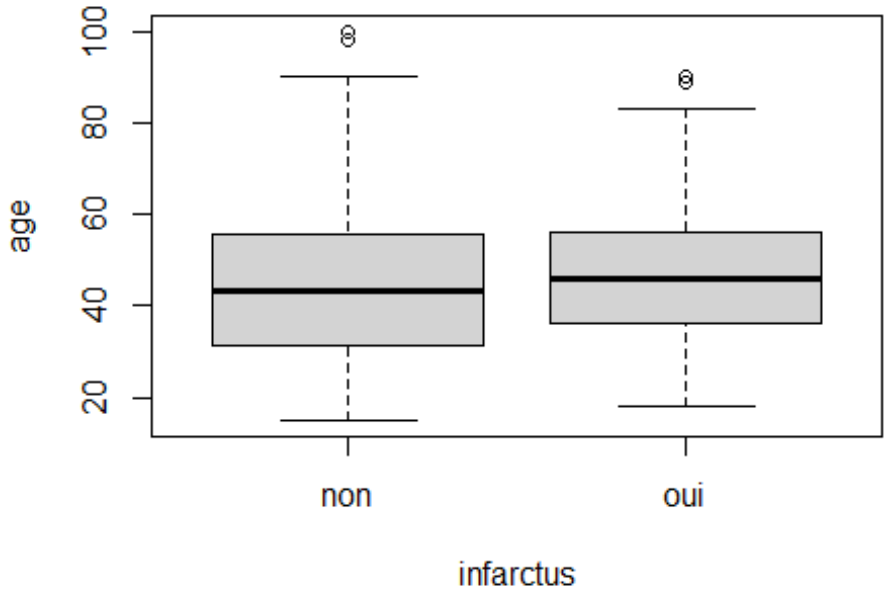
```
## hypertension
## non oui
## 290 159
```



III. Analyse Bivariée

A. Variables quantitatives VS Variable Infarctus

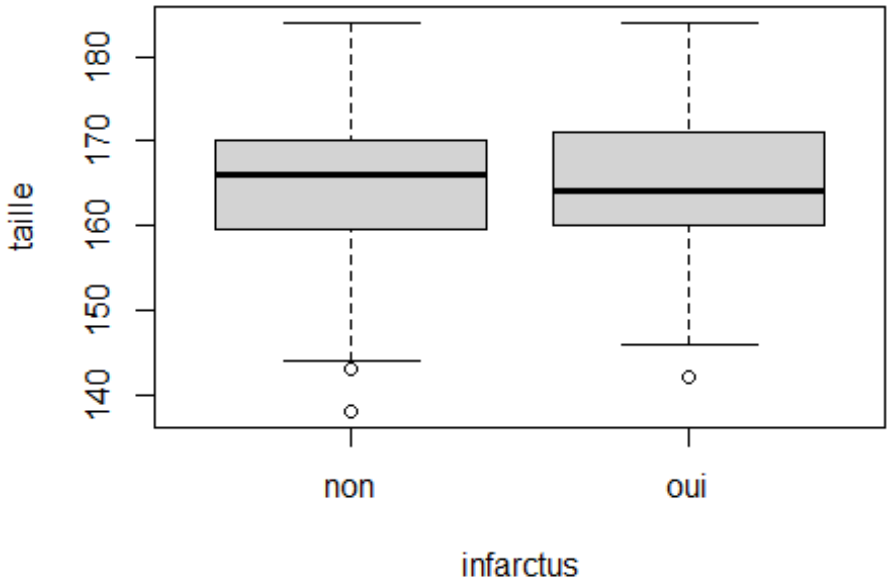
1. Age vs Infarctus




```
##
##  Welch Two Sample t-test
##
## data:  age by infarctus
## t = -1.2628, df = 328.11, p-value = 0.2076
## alternative hypothesis: true difference in means between group non and group oui is
not equal to 0
## 95 percent confidence interval:
##  -5.029210  1.096772
## sample estimates:
## mean in group non mean in group oui
##          44.96667          46.93289
```

Le diagramme en boîte compare la distribution des âges entre deux groupes : ceux ayant subi un infarctus ("oui") et ceux qui ne l'ont pas subi ("non"). Les médianes des deux groupes sont proches, suggérant que l'âge est réparti de manière similaire entre eux. On observe une légère tendance à un âge moyen plus élevé dans le groupe "oui" (46,93 ans) par rapport au groupe "non" (44,97 ans). Cependant, le **test de Welch** indique que cette différence n'est pas statistiquement significative (**p-value = 0,2076**), ce qui signifie qu'on ne peut pas rejeter l'hypothèse nulle d'une égalité des moyennes d'âge entre les deux groupes. L'intervalle de confiance à 95% (-5,03 ; 1,10) inclut zéro, renforçant l'idée qu'il n'y a pas de différence notable entre les deux groupes en termes d'âge.

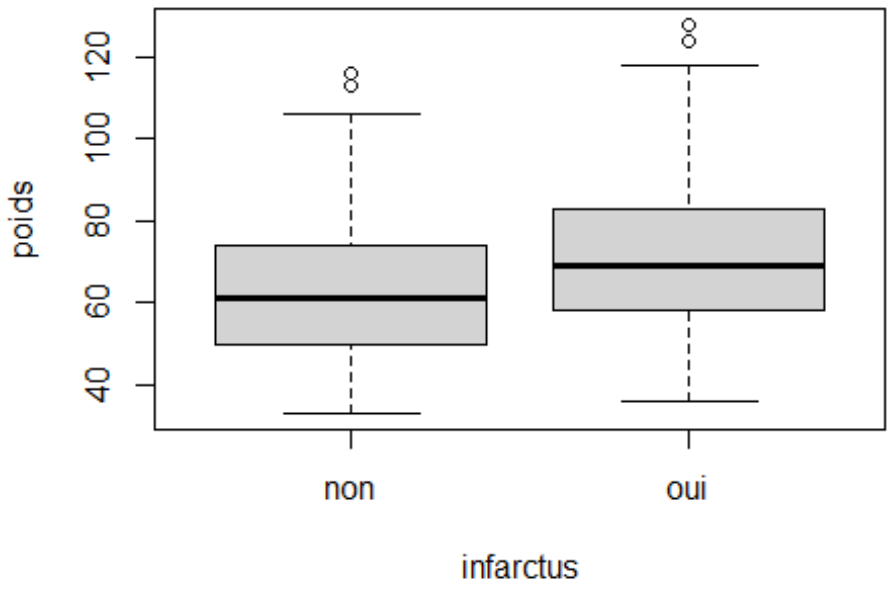
2. Taille vs Infarctus




```
##  Welch Two Sample t-test
##
## data:  taille by infarctus
## t = 0.71525, df = 295.25, p-value = 0.475
## alternative hypothesis: true difference in means between group non and group
oui is not equal to 0
## 95 percent confidence interval:
##  -1.018540  2.181583
## sample estimates:
## mean in group non mean in group oui
##          165.3533          164.7718
```

Le diagramme en boîte compare la distribution de la taille entre les individus ayant subi un infarctus ("oui") et ceux qui ne l'ont pas subi ("non"). Les médianes des deux groupes sont très proches, et les distributions sont similaires en termes d'étendue et de dispersion. La taille moyenne dans le groupe "non" est de **165,35 cm**, tandis que dans le groupe "oui", elle est légèrement inférieure à **164,77 cm**. Cependant, le **test de Welch** indique que cette différence de taille moyenne n’est pas statistiquement significative (**p-value = 0,475**). L’intervalle de confiance à 95 % (-1,02 ; 2,18) inclut zéro, ce qui signifie qu'on ne peut pas conclure à une différence réelle de taille entre les deux groupes. En résumé, la taille ne semble pas être un facteur discriminant entre les personnes ayant ou non subi un infarctus.

3. Poids vs Infarctus



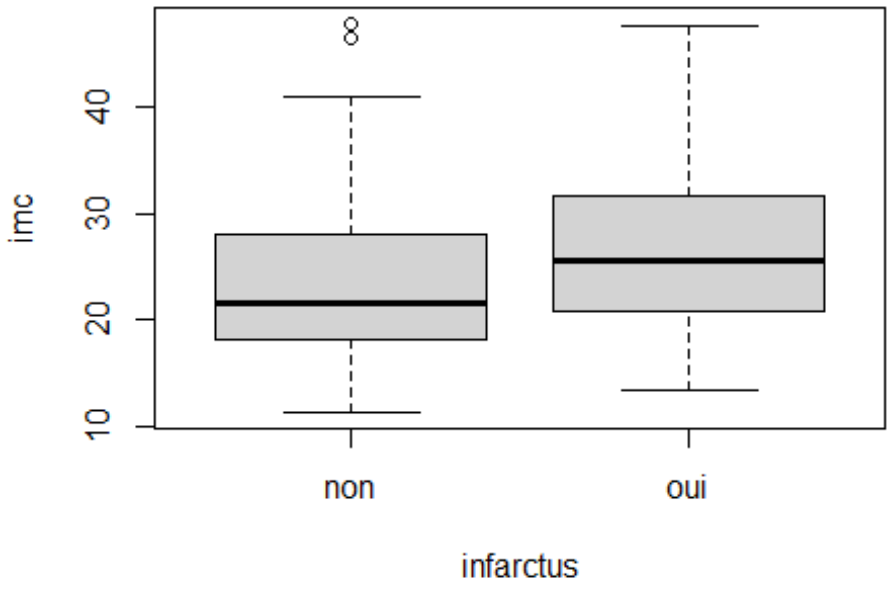

```
## data:  poids by infarctus
## t = -4.2048, df = 270.17, p-value = 3.556e-05
## alternative hypothesis: true difference in means between group non and group
oui is not equal to 0
## 95 percent confidence interval:
##  -11.310458  -4.096581
## sample estimates:
## mean in group non mean in group oui
##          63.51027          71.21379
```

Le diagramme en boîte montre la distribution du poids entre les individus ayant subi un infarctus et ceux qui ne l'ont pas subi. On constate que la médiane du poids est plus élevée chez les personnes ayant eu un infarctus, et que leur distribution est décalée vers des valeurs plus importantes. Les résultats statistiques confirment cette observation, avec une moyenne de 63,51 kg chez les non-infarctus contre 71,21 kg chez les infarctus, soit une différence moyenne de 7,70 kg. L'analyse statistique révèle une p-value extrêmement faible (3,556e-05), indiquant que cette différence est hautement significative. L'intervalle de confiance à 95 % [-11,31 ; -4,10] exclut zéro, renforçant l'idée d'une association entre un poids plus élevé et l'infarctus. Ces résultats suggèrent que les individus plus lourds ont un risque accru d'infarctus, ce qui est en accord avec les connaissances médicales établissant un lien entre l'obésité et les maladies cardiovasculaires.

4. IMC vs Infarctus

On a calculé la variable IMC.

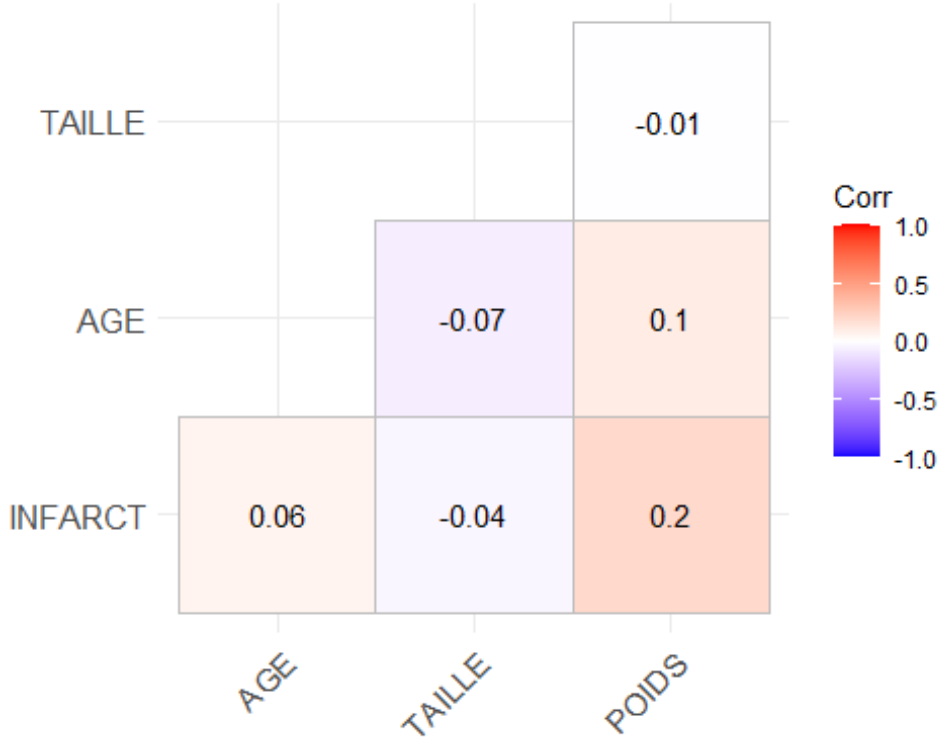
```
imc=poids/(taille/100)^2
```




```
##  Welch Two Sample t-test
##
## data:  imc by infarctus
## t = -4.1124, df = 280.29, p-value = 5.15e-05
## alternative hypothesis: true difference in means between group non and group
oui is not equal to 0
## 95 percent confidence interval:
##  -4.367171 -1.539737
## sample estimates:
## mean in group non mean in group oui
##          23.40426          26.35772
```

Le diagramme en boîte montre la distribution de l'imc entre les individus ayant subi un infarctus et ceux qui ne l'ont pas subi. On constate que la médiane de l'imc est plus élevée chez les personnes ayant eu un infarctus, et que leur distribution est décalée vers des valeurs plus importantes. Les résultats statistiques confirment cette observation, avec une moyenne de 26,404 chez les non-infarctus contre 26,358 chez les infarctus. L'analyse statistique révèle une p-value extrêmement faible (5,15e-05), indiquant que cette différence est hautement significative. L'intervalle de confiance à 95 % [-4,37 ; -1,54] exclut zéro, renforçant l'idée d'une association entre un imc plus élevé et l'infarctus.

5. Matrice de Corrélation



La matrice de corrélation indique que la variable INFARCT entretient des relations linéaires très faibles avec les autres variables analysées. Plus précisément, une légère corrélation positive est observée avec POIDS (0,2), suggérant qu'une augmentation du poids pourrait être associée à un risque légèrement plus élevé d'infarctus, bien que cette relation reste modeste. En revanche, la corrélation avec AGE est quasi nulle (0,06), indiquant que l'âge n'influence pratiquement pas la survenue d'infarctus dans ce contexte. La relation avec TAILLE est également négligeable (-0,04), montrant une absence de lien significatif entre la taille et l'infarctus. Globalement, ces résultats soulignent que les variables étudiées ici expliquent très peu la variabilité de l'infarctus, ce qui pourrait impliquer que d'autres facteurs non inclus dans cette analyse jouent un rôle plus déterminant.

B. Variables Qualitatives vs Infarctus

1. Contraceptif Oral vs Infarctus

```
##          contraceptif
## infarctus non oui
##       non 212  88
##       oui  37 112
```

```
## Total Observations in Table:  449
##
##
##                | contraceptif
##     infarctus  |       non |       oui | Row Total |
## -------------|-----------|-----------|-----------|
##           non |       212 |        88 |       300 |
##               |   166.370 |   133.630 |           |
##               |    12.515 |    15.581 |           |
##               |     0.707 |     0.293 |     0.668 |
##               |     0.851 |     0.440 |           |
##               |     0.472 |     0.196 |           |
```

```
## -------------|-----------|-----------|-----------|
##          oui |        37 |       112 |       149 |
##              |    82.630 |    66.370 |           |
##              |    25.198 |    31.372 |           |
##              |     0.248 |     0.752 |     0.332 |
##              |     0.149 |     0.560 |           |
##              |     0.082 |     0.249 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       249 |       200 |       449 |
##              |     0.555 |     0.445 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  84.66591     d.f. =  1     p =  3.532887e-20
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  82.8206     d.f. =  1     p =  8.984851e-20
```

Indicateur de spécificité:

```
contraceptif
## infarctus       non       oui
##       non 1.2742704 0.6585333
##       oui 0.4477777 1.6875168
```

La matrice de contingence et les tests statistiques révèlent une association significative entre l'utilisation de contraceptifs et la survenue d'infarctus. Le test du Chi² de Pearson ($\chi^2 = 84{,}67$, $p < 0{,}05$) et sa version avec correction de Yates ($\chi^2 = 82{,}82$, $p < 0{,}05$) confirment un lien statistiquement robuste, rejetant l'hypothèse d'indépendance entre les variables. Les indicateurs de spécificité renforcent cette observation : les utilisatrices de contraceptifs (« oui ») présentent une spécificité élevée pour l'infarctus (1,69), indiquant une fréquence nettement supérieure à celle attendue sous l'hypothèse d'absence de lien. À l'inverse, les non-utilisatrices (« non ») sont associées à une moindre occurrence d'infarctus (spécificité de 0,45 pour « oui »). Par ailleurs, 75 % des cas d'infarctus (112/149) concernent des utilisatrices de contraceptifs, contre seulement 29 % (88/300) chez les non-cas, soulignant un déséquilibre marqué. Ces résultats suggèrent que l'utilisation de contraceptifs pourrait être un facteur de risque associé à l'infarctus dans cette population, bien que des études complémentaires soient nécessaires pour explorer les mécanismes sous-jacents.

2. Tabac vs Infarctus

```
##          tabac
## infarctus nonfumeuse actuelle ancienne
##       non        181       75       44
##       oui         34       60       55
```

```
## Total Observations in Table:  449
##
##
##                | tabac
##    infarctus | nonfumeuse |   actuelle |   ancienne |  Row Total |
## -------------|------------|------------|------------|------------|
##          non |        181 |         75 |         44 |        300 |
##              |    143.653 |     90.200 |     66.147 |            |
##              |      9.710 |      2.562 |      7.415 |            |
##              |      0.603 |      0.250 |      0.147 |      0.668 |
##              |      0.842 |      0.556 |      0.444 |            |
##              |      0.403 |      0.167 |      0.098 |            |
## -------------|------------|------------|------------|------------|
##          oui |         34 |         60 |         55 |        149 |
##              |     71.347 |     44.800 |     32.853 |            |
##              |     19.550 |      5.157 |     14.930 |            |
##              |      0.228 |      0.403 |      0.369 |      0.332 |
##              |      0.158 |      0.444 |      0.556 |            |
##              |      0.076 |      0.134 |      0.122 |            |
## -------------|------------|------------|------------|------------|
## Column Total |        215 |        135 |         99 |        449 |
##              |      0.479 |      0.301 |      0.220 |            |
## -------------|------------|------------|------------|------------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  59.32361     d.f. =  2     p =  1.312328e-13
```

Indicateur de spécificité :

```
##          tabac
## infarctus nonfumeuse  actuelle  ancienne
##       non  1.2599845 0.8314815 0.6651852
##       oui  0.4765413 1.3392990 1.6741238
```

La matrice de contingence et les tests statistiques mettent en évidence une association significative entre le statut de consommation du tabac et la survenue d'infarctus. Le test du Chi² de Pearson ($\chi^2 = 59{,}32$, ddl = 2, $p < 0{,}05$) rejette fortement l'hypothèse d'indépendance entre ces variables. Les indicateurs de spécificité révèlent des profils distincts : les non-fumeuses auront tendance à ne pas souffrir d'infarctus (spécificité = 1,26). À l'inverse, les fumeuses actuelles et anciennes présentent des spécificités élevées pour l'infarctus (1,34 et 1,67 respectivement), indiquant une fréquence accrue par rapport aux attentes sous l'hypothèse d'absence de lien. Parmi les cas d'infarctus, 40,3 % sont des fumeuses actuelles et 36,9 % d'anciennes fumeuses, contre seulement 25 % et 14,7 % chez les non-cas. Les anciennes fumeuses affichent la spécificité la plus marquée, suggérant un impact durable du tabagisme même après l'arrêt. Ces résultats confirment que le tabagisme—actuel ou passé—est un facteur associé à un risque accru d'infarctus dans cette population.

3. Antécédent familiaux vs Infarctus

```
##          antecedent
## infarctus non oui
```

```
##       non 265  31
##       oui 124  22
```

```
## Total Observations in Table:  442
##
##
##                | antecedent
##    infarctus |       non |       oui | Row Total |
## -------------|-----------|-----------|-----------|
##          non |       265 |        31 |       296 |
##              |   260.507 |    35.493 |           |
##              |     0.077 |     0.569 |           |
##              |     0.895 |     0.105 |     0.670 |
##              |     0.681 |     0.585 |           |
##              |     0.600 |     0.070 |           |
## -------------|-----------|-----------|-----------|
##          oui |       124 |        22 |       146 |
##              |   128.493 |    17.507 |           |
##              |     0.157 |     1.153 |           |
##              |     0.849 |     0.151 |     0.330 |
##              |     0.319 |     0.415 |           |
##              |     0.281 |     0.050 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       389 |        53 |       442 |
##              |     0.880 |     0.120 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  1.956639     d.f. =  1     p =  0.1618732
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  1.545403     d.f. =  1     p =  0.2138154
```

Indicateur de spécificité :

```
##          antecedent
## infarctus       non       oui
##       non 1.0172480 0.8734064
##       oui 0.9650315 1.2566555
```

Le test du Chi² ($\chi^2 = 1{,}96$, ddl = 1, p = 0,16) et sa version avec correction de Yates ($\chi^2 = 1{,}55$, p = 0,21) ne permettent pas de rejeter l'hypothèse d'indépendance entre les deux variables, les résultats n'étant pas statistiquement significatifs au seuil conventionnel de 5 %. Les indicateurs de spécificité reflètent cette absence de lien marqué : les non-cas d'infarctus sont très légèrement associés à l'absence d'antécédents (spécificité = 1,02), tandis que les cas d'infarctus montrent une spécificité modérément élevée pour les antécédents (1,26). Cependant, ces écarts restent faibles et cohérents avec des variations aléatoires, comme le confirment les résidus standardisés proches de zéro. Une exploration plus approfondie, intégrant

éventuellement d'autres variables ou une taille d'échantillon plus importante, pourrait être nécessaire pour clarifier ce lien.

4. Hypertension vs Infarctus

```
##          hypertension
## infarctus non oui
##       non 205  95
##       oui  85  64

## Total Observations in Table:  449
##
##
##              | hypertension
##    infarctus |       non |       oui | Row Total |
## -------------|-----------|-----------|-----------|
##          non |       205 |        95 |       300 |
##              |   193.764 |   106.236 |           |
##              |     0.652 |     1.188 |           |
##              |     0.683 |     0.317 |     0.668 |
##              |     0.707 |     0.597 |           |
##              |     0.457 |     0.212 |           |
## -------------|-----------|-----------|-----------|
##          oui |        85 |        64 |       149 |
##              |    96.236 |    52.764 |           |
##              |     1.312 |     2.393 |           |
##              |     0.570 |     0.430 |     0.332 |
##              |     0.293 |     0.403 |           |
##              |     0.189 |     0.143 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       290 |       159 |       449 |
##              |     0.646 |     0.354 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  5.544547     d.f. =  1     p =  0.01853836
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  5.062067     d.f. =  1     p =  0.02445506
```

Indicateur de spécificité :

```
##          hypertension
## infarctus       non       oui
##       non 1.0579885 0.8942348
##       oui 0.8832446 1.2129501
```

Le test du Chi² ($\chi^2 = 5,54$, ddl = 1, p = 0,019) et sa version avec correction de Yates ($\chi^2 = 5,06$, p = 0,024) rejettent l'hypothèse d'indépendance entre ces variables au seuil de 5 %. Les indicateurs de spécificité soulignent une tendance claire : les cas d'infarctus sont associés à une spécificité plus élevée pour l'hypertension (1,21), tandis que les non-cas montrent une spécificité légèrement supérieure pour l'absence d'hypertension (1,06). Parmi les cas d'infarctus, 43 % (64/149) présentent une hypertension, contre 31,7 % (95/300) chez les non-cas, ce qui traduit une surreprésentation. Ces résultats suggèrent que l'hypertension pourrait être un facteur de risque modéré pour l'infarctus dans cette population.

IV. Analyse Multivariée

A. ACP sur des données quantitatives

- On clone le dataframe initial

```
df_quanti <- data.frame(df$INFARCT, df$CO, df$TABAC, df$AGE, imc, df$ATCD, df$HTA)
```

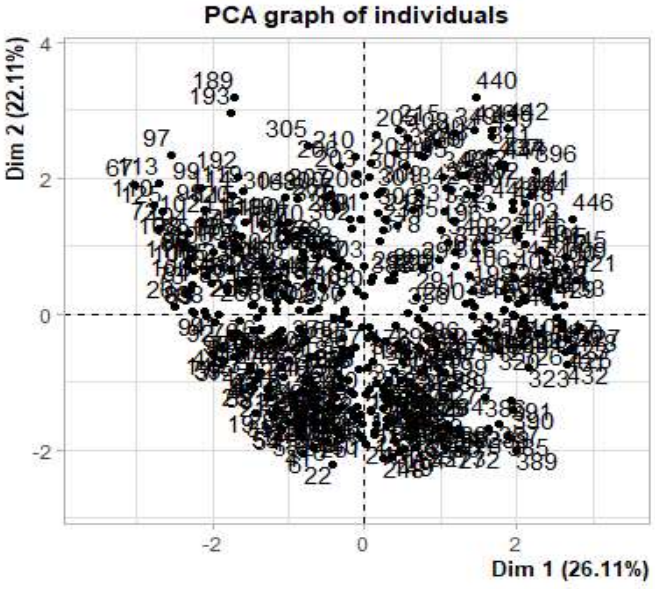
- On supprime les valeurs nulles et on normalise nos données

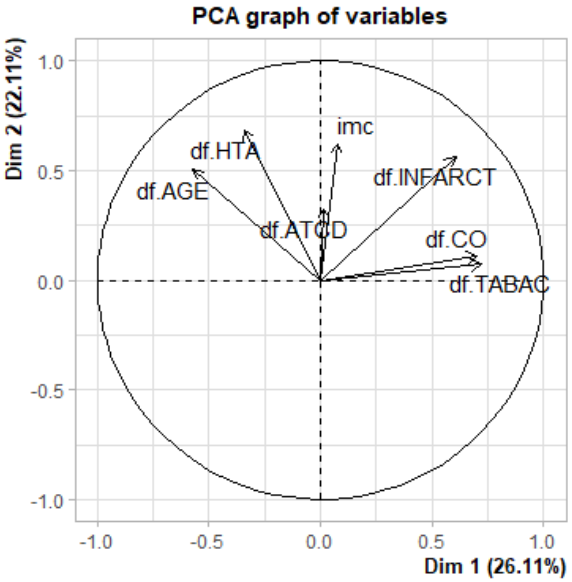
```
df_quanti <- na.omit(df_quanti)

df_quanti <- scale(df_quanti)
```

- On fait l'ACP

```
acp_v1 <- PCA(df_quanti)
```





Les deux axes, Dim 1 (26,11%) et Dim 2 (22,11%), expliquent ensemble environ 48,22% de la variance totale des données. Les variables fortement corrélées sont proches les unes des autres et orientées dans la même direction. Ici, on observe que **df.TABAC** et **df.CO** sont fortement corrélées avec la première

dimension (Dim 1), tandis que **df.HTA** et **imc** contribuent davantage à la deuxième dimension (Dim 2). La variable **df.INFARCT** est bien représentée sur la première dimension (Dim 1), avec une orientation proche de celle de **df.TABAC** et **df.CO**. Cela suggère une corrélation positive entre l'infarctus et ces variables, ce qui signifie que des niveaux élevés de tabagisme et de consommation de contraceptif oral sont associés à une plus grande probabilité d'infarctus. En revanche, **df.INFARCT** est peu projetée sur la deuxième dimension (Dim 2), indiquant qu'elle n'est pas fortement liée aux variables comme **df.HTA** et **df.AGE** qui influencent davantage cette composante.

B. MCA sur variables qualitatives

- Recodage des variables en variables qualitatives

Les variables déjà qualitatives sont juste mises dans de nouvelles variables précédées de la lettre 'q' pour qualitative.

```
qinfarctus <- infarctus
qcontraceptif <- contraceptif
qtabac <- tabac
qantecedent <- antecedent
qhypertension <- hypertension
```

Les variables quantitatives sont regroupées en fonction des quartiles pour faire des goupes.

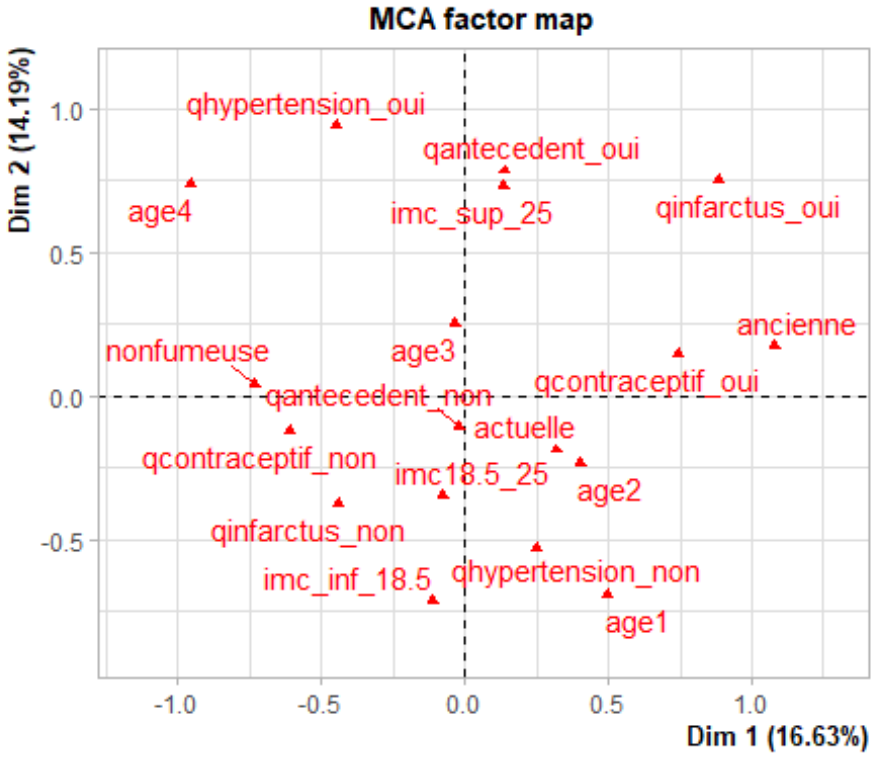
```
qage <- cut(age, breaks =c(14,33,44,56,100), labels=c("age1", "age2", "age3", "age4"))
qpoids <- cut(poids, breaks =c(30,51,64,79,130), labels=c("poids1", "poids2", "poids3"
, "poids4"))
qtaille <- cut(taille, breaks =c(130,160,166,171,184), labels=c("taille1", "taille2",
"taille3", "taille4"))
qimc <- cut(imc, breaks =c(0,18.5,25,50), labels=c("imc_inf_18.5", "imc18.5_25", "imc_
sup_25"))
```

- On crée un dataframe composé de toutes ces variables qualitatives

```
df_quali <- data.frame(qinfarctus, qcontraceptif, qantecedent, qhypertension, qtabac,
qage, qimc)
df_quali <- na.omit(df_quali)
```

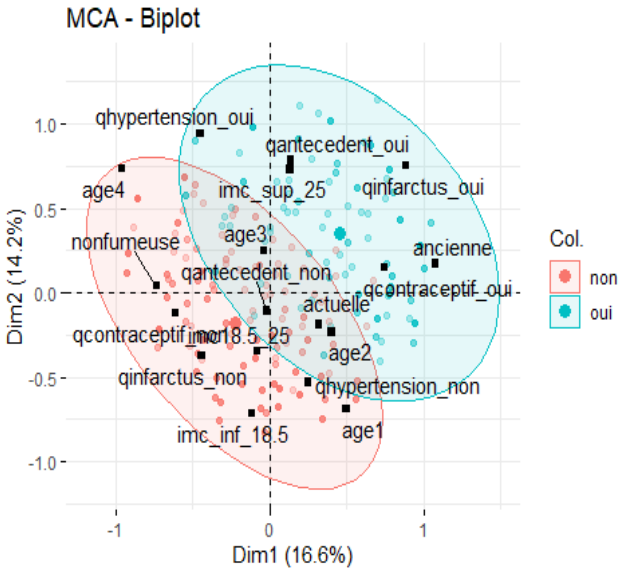
- On applique le MCA

```
mca_v1 <- MCA(df_quali)
```



- On crée un biplot pour représenter les groupes de modalités prônes à un infarctus et celles non prônes.

```
fviz_mca_biplot(mca_v1,
                repel = TRUE,
                geom.ind = c("point"),
                col.ind = df_quali$qinfarctus,  # Couleur des individues
                # col.ind = as.factor(groupe.kmeans),  # Couleur des individues
                alpha.ind = 0.3,
                geom.var = c("point", "text"),
                col.var = "black", # Couleur des variables
                # arrows = c(FALSE, TRUE),
                # pointsize = 1,
                shape.var = 15,
                size.var = 6,
                addEllipses = TRUE
)
```



Les axes **Dim 1 (16,6%)** et **Dim 2 (14,2%)** expliquent ensemble environ **30,8%** de l'inertie totale, ce qui signifie que ces deux dimensions résument une part significative de la variabilité des données.

- **Zone bleue ("oui")** : Cette zone regroupe les individus ayant eu un infarctus (**qinfarctus_oui**). On y trouve des caractéristiques associées, comme **qhypertension_oui**, **qantecedent_oui** (antécédents familiaux positifs), **qcontraceptif_oui** (contraceptifs oraux utilisés), et un **IMC supérieur à 25**. On note aussi la présence de l'attribut **"ancienne"**, ce qui suggère que les anciennes fumeuses ont plus de probabilité d'avoir un infarctus. Il n'y a pas de modalités **d'âge** dans cette zone, donc on aura tendance à penser que l'âge n'influence pas le risque d'infarctus.
- **Zone rouge ("non")** : Cette zone regroupe les individus n'ayant pas eu d'infarctus (**qinfarctus_non**). Elle est associée à **qhypertension_non**, **nonfumeuse**, **qcontraceptif_non**, et un **IMC inférieur à 18,5**. C Donc les personnes sans hypertension, ne prenant pas de contraceptif oral et non fumeuses, ont plus de chance de ne pas souffrir d'un infarctus.
- **Zone frontière** : Certaines modalités se retrouvent à la frontière du bleu et du rouge. A l'exemple de **qantecedent_non**, **age2**, **age3** et **actuelle**. On pourra donc conclure que le fait de ne pas d'antécédents familiaux, d'être d'âge moyen (33-56ans) et consommé activement du tabac n'influençait pas les risques d'infarctus.

V. Modelisation

A. Préparation à la modèlisation

- Chargement de la librairie caret

```
library(caret)
```

caret (Classification And Regression Training) est une bibliothèque R très utilisée pour l'entraînement et l'évaluation de modèles de machine learning.

- Division de l'ensemble d'entrainement et de l'ensemble de test

```
set.seed(123)
taux_train <- 0.8
indices <- sample(nrow(df_quali), nrow(df_quali)*taux_train)
df_train <- df_quali[indices,]
df_test <- df_quali[-indices,]
```

On fixe la graine aléatoire. Ensuite on choisit un taux de 0.8, c'est-à-dire que 80% du dataset est attribué à l'entrainement et 20% au test.

B. Régression Logistique

```
logistic <- train(qinfarctus~., data=df_train,
                  method="glm",
                  trControl = trainControl(method="repeatedcv",
                                           number=10,
                                           repeats=3),
                  family="binomial")
```

On utilise la méthode train de la librairie caret. `qinfarctus~.` Nous permet de préciser que la variable cible est qinfarctus et les variables indépentandes sont toutes les autres variables du dataframe. `data=df_train` permet de spécifier quel dataframe est à utiliser. `method="glm"` précise quel modèle de classification ou de régression utiliser. Ici on choisi glm, c'est-à-dire Generalized Linear Model ce qui nous

permet de faire la régression logistique. `family="binomial"` permet de préciser qu'il s'agira d'une classification binaire.

Enfin, **`trainControl`**. `method="repeatedcv"` Utilise une validation croisée répétée. `number=10` Effectue une validation croisée à **10 folds** (les données sont divisées en 10 sous-ensembles). `repeats=3` Répète la validation croisée **3 fois** pour améliorer la robustesse de l'évaluation.

```
summary(logistic)
```

```
## Call:
## NULL
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.8391     0.6158  -7.858 3.90e-15 ***
## qcontraceptifoui   2.1876     0.3076   7.113 1.14e-12 ***
## qantecedentoui     0.1782     0.4282   0.416   0.6773
## qhypertensionoui   0.5421     0.3259   1.663   0.0963 .
## qtabacactuelle     1.8982     0.3655   5.194 2.06e-07 ***
## qtabacancienne     2.4231     0.4228   5.732 9.95e-09 ***
## qageage2           1.0444     0.4165   2.507   0.0122 *
## qageage3           1.2948     0.4265   3.036   0.0024 **
## qageage4           2.0735     0.5130   4.042 5.30e-05 ***
## qimcimc18.5_25     0.2135     0.3918   0.545   0.5858
## qimcimc_sup_25     0.8036     0.3828   2.099   0.0358 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 447.39  on 343  degrees of freedom
## Residual deviance: 315.14  on 333  degrees of freedom
## AIC: 337.14
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(logistic$finalModel))
```

```
##      (Intercept) qcontraceptifoui   qantecedentoui qhypertensionoui
##      0.007914329      8.914219556      1.195046927      1.719567917
##   qtabacactuelle   qtabacancienne         qageage2         qageage3
##      6.674064218     11.281018027      2.841669049      3.650153341
##         qageage4   qimcimc18.5_25   qimcimc_sup_25
##      7.952281007      1.238005584      2.233615669
```

```
library(vip)
```

```
## Warning: package 'vip' was built under R version 4.4.3
```

```
##
## Attaching package: 'vip'
```

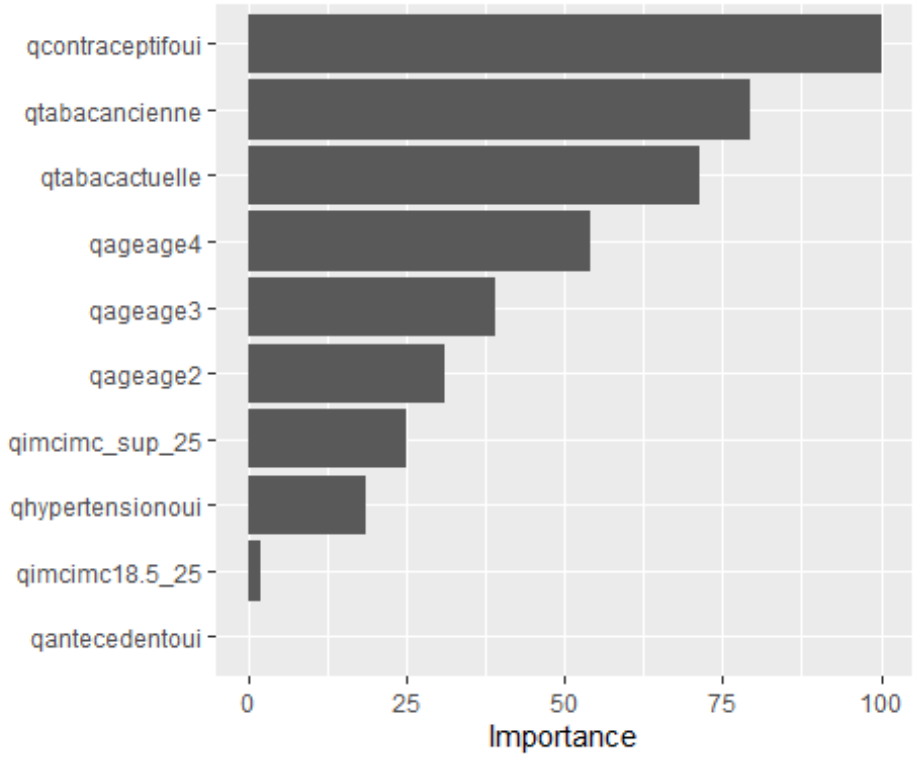
```
## The following object is masked from 'package:utils':
##
##     vi
```

```
importance_logit <- vip(logistic, num_features =20)
importance_logit$data
```

```
## # A tibble: 10 × 2
##    Variable         Importance
```

```
##    <chr>                <dbl>
##  1 qcontraceptifoui     100
##  2 qtabacancienne        79.4
##  3 qtabacactuelle        71.3
##  4 qageage4              54.1
##  5 qageage3              39.1
##  6 qageage2              31.2
##  7 qimcimc_sup_25        25.1
##  8 qhypertensionoui      18.6
##  9 qimcimc18.5_25         1.92
## 10 qantecedentoui         0
```

```
plot(importance_logit)
```




```
predictions_logit <- predict(logistic, newdata=df_test)
confusionMatrix(predictions_logit, df_test$qinfarctus)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction non oui
##        non  65   8
##        oui   1  13
##
##                Accuracy : 0.8966
##                  95% CI : (0.8127, 0.9516)
##     No Information Rate : 0.7586
##     P-Value [Acc > NIR] : 0.0009444
##
##                   Kappa : 0.6813
##
##  Mcnemar's Test P-Value : 0.0455003
##
##             Sensitivity : 0.9848
```

```
##             Specificity : 0.6190
##          Pos Pred Value : 0.8904
##          Neg Pred Value : 0.9286
##              Prevalence : 0.7586
##          Detection Rate : 0.7471
##    Detection Prevalence : 0.8391
##       Balanced Accuracy : 0.8019
##
##        'Positive' Class : non
```

Les résultats des coefficients exponentiés (OR=Odd Ratios) indiquent le facteur d'augmentation du risque d'infarctus associé à chaque variable :

- **Utilisation de contraceptifs (qcontraceptifoui)** : Augmente fortement le risque d'infarctus (OR = **8.91**).
- **Consommation de tabac** : Les fumeurs actuels (**qtabacactuelle**, OR = **6.67**) et anciens fumeurs (**qtabacancienne**, OR = **11.28**) ont un risque significativement accru, les anciens fumeurs étant encore plus à risque.
- **Âge** : Comparé au groupe le plus jeune, le risque d'infarctus augmente progressivement avec l'âge : **qageage2 (OR = 2.84), qageage3 (OR = 3.65), qageage4 (OR = 7.95)**.
- **IMC** : Un IMC supérieur à 25 (**qimcimc_sup_25**, OR = **2.23**) augmente le risque d'infarctus, alors qu'un IMC entre 18,5 et 25 (qimcimc18.5_25) a un effet moindre (OR = **1.24**).
- **Hypertension (qhypertensionoui)** : Augmente le risque d'infarctus (OR = **1.72**), bien que l'effet soit moins marqué que d'autres facteurs.
- **Antécédents familiaux (qantecedentoui)** : Ont un effet faible sur le risque d'infarctus (OR = **1.19**), ce qui suggère que des facteurs liés au mode de vie pourraient avoir une influence plus forte.

L'analyse de l'importance des variables montre que les **contraceptifs (100 %)**, le **tabagisme passé (79.4 %) et actuel (71.3 %)** sont les principaux facteurs de risque d'infarctus. L'âge est également déterminant, en particulier pour le groupe des **65 ans et plus (qageage4, 54.1 %)**. L'IMC élevé et l'hypertension jouent un rôle modéré, tandis que les antécédents familiaux ont une importance très faible (0 %).

Le modèle présente une **précision globale de 89.66 %**, avec une bonne capacité à identifier les personnes **sans infarctus (sensibilité : 98.48 %)**, mais une capacité plus limitée à détecter correctement les cas d'infarctus (spécificité : 61.90 %). L'indice Kappa (0.6813) indique une bonne concordance entre les prédictions et la réalité.

C. Random Forest

```
rf <- train(qinfarctus~., data=df_train,
                  method="rf",
                  trControl = trainControl(method="repeatedcv",
                                           number=10,
                                           repeats=3),
                  ntree=500)
```

`Ici, method="rf"` indique que la méthode de classification choisi est le random forest, avec `ntree=500` qui définit le nombre d'arbres à **500** dans la forêt aléatoire, ce qui aide à améliorer la robustesse des prédictions.

```
summary(rf)
```

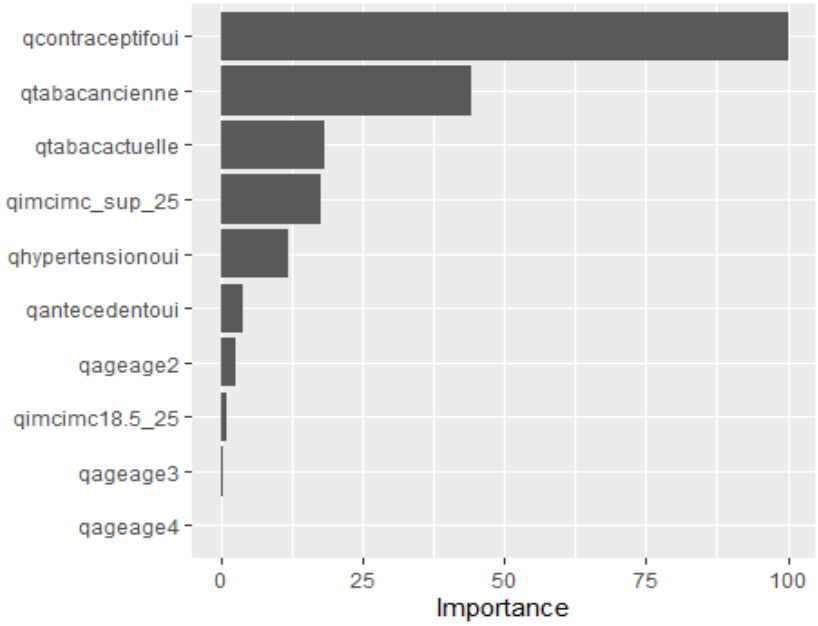
```
##                 Length Class      Mode
## call               5   -none-     call
```

```
## type                1   -none-     character
## predicted         344   factor     numeric
## err.rate         1500   -none-     numeric
## confusion           6   -none-     numeric
## votes             688   matrix     numeric
## oob.times         344   -none-     numeric
## classes             2   -none-     character
## importance         10   -none-     numeric
## importanceSD        0   -none-     NULL
## localImportance     0   -none-     NULL
## proximity           0   -none-     NULL
## ntree               1   -none-     numeric
## mtry                1   -none-     numeric
## forest             14   -none-     list
## y                 344   factor     numeric
## test                0   -none-     NULL
## inbag               0   -none-     NULL
## xNames             10   -none-     character
## problemType         1   -none-     character
## tuneValue           1   data.frame list
## obsLevels           2   -none-     character
## param               1   -none-     list
```

```
importance_rf <- vip(rf, num_features =20)
importance_rf$data
```

```
## # A tibble: 10 × 2
##    Variable         Importance
##    <chr>                 <dbl>
##  1 qcontraceptifoui    100
##  2 qtabacancienne       44.4
##  3 qtabacactuelle       18.4
##  4 qimcimc_sup_25       17.6
##  5 qhypertensionoui     11.9
##  6 qantecedentoui        4.01
##  7 qageage2              2.71
##  8 qimcimc18.5_25        1.00
##  9 qageage3              0.572
## 10 qageage4              0
```

```
plot(importance_rf)
```



```r
predictions_rf <- predict(rf, newdata=df_test)
confusionMatrix(predictions_rf, df_test$qinfarctus)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction non oui
##        non  63  10
##        oui   3  11
##
##                Accuracy : 0.8506
##                  95% CI : (0.758, 0.918)
##     No Information Rate : 0.7586
##     P-Value [Acc > NIR] : 0.02571
##
##                   Kappa : 0.5397
##
##  Mcnemar's Test P-Value : 0.09609
##
##             Sensitivity : 0.9545
##             Specificity : 0.5238
##          Pos Pred Value : 0.8630
##          Neg Pred Value : 0.7857
##              Prevalence : 0.7586
##          Detection Rate : 0.7241
##    Detection Prevalence : 0.8391
##       Balanced Accuracy : 0.7392
##
##        'Positive' Class : non
```

L'analyse de l'importance des variables montre que :

- **La prise de contraceptifs (qcontraceptifoui)** est le facteur le plus déterminant, avec une importance normalisée à 100.
- **Le tabagisme, qu'il soit actuel (qtabacactuelle) ou ancien (qtabacancienne)**, joue un rôle clé, avec une forte contribution à la prédiction du risque d'infarctus.
- **L'IMC supérieur à 25 (qimcimc_sup_25) et l'hypertension (qhypertensionoui)** influencent également le risque, bien que dans une moindre mesure.
- **Les antécédents familiaux (qantecedentoui) ont une faible importance relative**, suggérant qu'ils sont moins prédictifs dans ce modèle que d'autres facteurs comportementaux comme le tabagisme.
- **L'âge (qageage2, qageage3, qageage4) a une importance très faible**, voire nulle pour les catégories les plus âgées, ce qui peut indiquer une interaction avec d'autres variables comme l'hypertension ou l'IMC.

La matrice de confusion permet d'évaluer la performance du modèle Random Forest sur l'ensemble de test :

- **Précision globale (Accuracy) : 85,06 %**, indiquant que le modèle fait de bonnes prédictions générales.
- **Sensibilité : 95,45 %**, ce qui signifie que le modèle détecte correctement les individus sans infarctus dans 95,45 % des cas.
- **Spécificité : 52,38 %**, ce qui montre que le modèle a plus de mal à identifier correctement les personnes ayant eu un infarctus (beaucoup de faux négatifs).

D. SVM

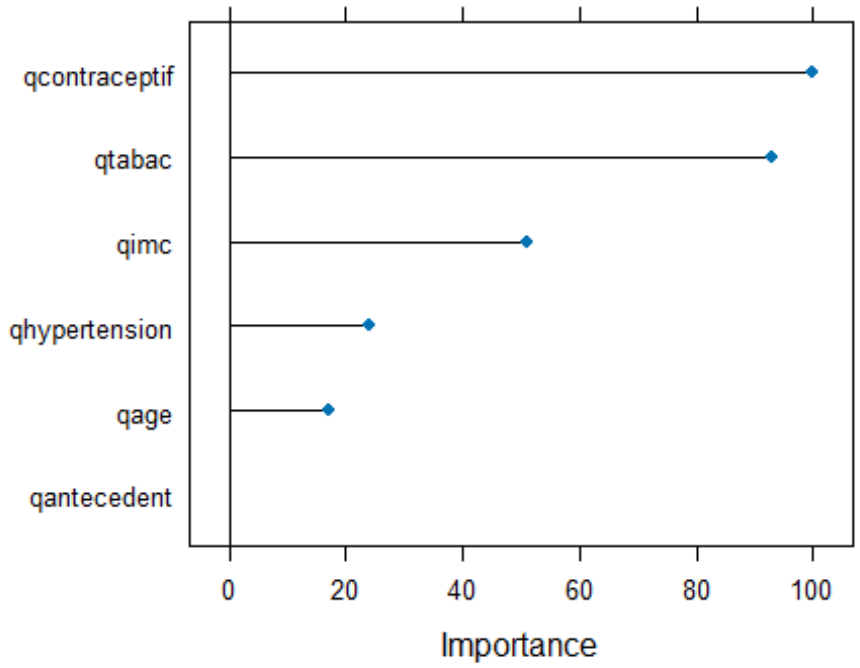
```
svm <- train(qinfarctus~., data=df_train,
                    method="svmRadial",
                    trControl = trainControl(method="repeatedcv",
                                             number=10,
                                             repeats=3),)
```

Finalement, `method="svmRadial"` permet de préciser que le modèle de classification ici est le svmradial.

```
summary(svm)
```

```
## Length  Class   Mode
##      1   ksvm     S4
```

```
importance_svm <- varImp(svm)
plot(importance_svm)
```




```
predictions_svm <- predict(svm, newdata=df_test)
confusionMatrix(predictions_svm, df_test$qinfarctus)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction non oui
##        non  63  10
##        oui   3  11
##
##                Accuracy : 0.8506
##                  95% CI : (0.758, 0.918)
##     No Information Rate : 0.7586
##     P-Value [Acc > NIR] : 0.02571
##
##                   Kappa : 0.5397
```

```
##
##  Mcnemar's Test P-Value : 0.09609
##
##             Sensitivity : 0.9545
##             Specificity : 0.5238
##          Pos Pred Value : 0.8630
##          Neg Pred Value : 0.7857
##              Prevalence : 0.7586
##          Detection Rate : 0.7241
##    Detection Prevalence : 0.8391
##       Balanced Accuracy : 0.7392
##
##        'Positive' Class : non
```

L'analyse de l'importance des variables montre quelles caractéristiques influencent le plus les prédictions du modèle SVM.

- **qcontraceptif** : La variable ayant l'importance la plus élevée. Cela pourrait indiquer une corrélation entre l'utilisation de contraceptifs et le risque d'infarctus dans l'échantillon étudié.
- **qtabac** : Le tabagisme est un facteur bien connu de maladies cardiovasculaires. Son importance élevée dans le modèle confirme cette relation.
- **qimc** : L'indice de masse corporelle (IMC) joue un rôle important dans le risque cardiovasculaire, notamment via l'obésité et le surpoids.
- **qhypertension** : L'hypertension est un facteur de risque majeur d'infarctus, ce qui justifie sa présence parmi les variables influentes.
- **qage** : L'âge est un facteur clé, bien qu'il ait une importance relativement plus faible que les autres facteurs cités précédemment.
- **qantecedent** : La présence d'antécédents médicaux a un impact moindre selon le modèle, mais reste un facteur à considérer.

L'évaluation du modèle SVM sur l'ensemble de test donne les résultats suivants :

- **Précision globale (Accuracy) : 85,06 %**, indiquant une bonne performance générale.
- **Sensibilité : 95,45 %**, signifiant que le modèle détecte correctement les personnes sans infarctus dans 95,45 % des cas.
- **Spécificité : 52,38 %**, ce qui montre que le modèle a des difficultés à identifier correctement les individus ayant eu un infarctus.

E. Comparaison des modèles

Modèle	Accuracy	Sensibilité	Spécificité	PPV	NPV
Régression logistique	89,66 %	98,48 %	61,90 %	89,04 %	92,86 %
Random Forest	85,06 %	95,45 %	52,38 %	86,30 %	78,57 %
SVM	85,06 %	95,45 %	52,38 %	86,30 %	78,57 %

Observations :

- La **régression logistique** a **la meilleure performance globale** avec une précision de **89,66 %** et une **spécificité plus élevée (61,90 %)** que les autres modèles.
- **Random Forest et SVM ont exactement les mêmes performances**, avec une précision inférieure à la régression logistique et une spécificité plus faible.

- La **sensibilité est élevée dans tous les modèles**, indiquant une très bonne capacité à identifier les individus sans infarctus.
- La **faible spécificité (52,38 % pour RF et SVM)** signifie que ces modèles ont tendance à classer à tort certaines personnes comme "sans infarctus".

Conclusion

Cette étude visait à analyser les facteurs influençant le risque d'infarctus à l'aide de plusieurs modèles d'apprentissage automatique : **régression logistique, Random Forest et Support Vector Machine (SVM)**. L'objectif était d'évaluer leurs performances et de déterminer les variables les plus influentes dans la prédiction du risque d'infarctus.

Les résultats obtenus montrent que **la prise de contraceptifs oraux, le tabagisme (actuel ou passé), l'hypertension et l'indice de masse corporelle (IMC) sont des facteurs déterminants**. En particulier, la prise de contraceptifs et le tabagisme semblent jouer un rôle clé dans l'augmentation du risque d'infarctus.

En termes de performance des modèles :

- **La régression logistique** offre une précision élevée (**89,66 %**) avec une bonne capacité à détecter les personnes à risque, mais sa capacité à identifier correctement les infarctus reste modérée.
- **Le Random Forest et le SVM** ont des performances similaires (**85,06 % de précision**), avec une meilleure sensibilité mais une spécificité plus faible.

Pour améliorer la détection des cas d'infarctus, plusieurs pistes peuvent être envisagées :

- **Optimisation des modèles** : ajustement des hyperparamètres du SVM et du Random Forest pour améliorer la spécificité.
- **Enrichissement des données** : prise en compte de nouvelles variables cliniques ou comportementales.
- **Utilisation d'autres types d'encodage en variables qualitatives que la segmentation par quartiles**

En conclusion, **l'identification précoce des facteurs de risque reste un enjeu majeur** pour la prévention des infarctus, et les modèles étudiés peuvent contribuer à améliorer le dépistage et la prise en charge des patients à risque.