Pour le 07/05/2024

Traitement Numériques de Données (TND)

Binôme 3

Rapport du jeu de données : Haberman

Réalisé par:

- *Camélia MAZOUZ*
- *Rémy XU*
- *Sami KRIM*

2023/2024

Table des matières

Table de figures

1.Introduction

Dans le cadre du cours de Traitement Numérique de Données, nous avons été chargés de rédiger un rapport détaillant nos observations et nos commentaires après avoir analysé le jeu de données **Haberman**. Cette tâche implique également de décrire en détail les méthodes utilisées, ainsi que d'analyser et de commenter les résultats obtenus.

2.Description du jeu de données

2.1.Vue globale

Les données nous ont été transmises via 2 fichiers:

- haberman.data : qui contient les données dont les colonnes sont séparées par une virgule “,” et les lignes par un retour à la ligne.
- haberman.names : qui contient plus d’informations sur la nature, la provenance des données et les noms des différentes colonnes du jeu de données.

2.2.Lecture du jeu de données

Le jeu de données **Haberman** ne contient aucune valeur manquante (**0**) et est composé de **306** instances définies par les **4** variables suivantes :

- **Age of patient at time of operation (numérique, entier) :** Représente l'âge du patient au moment de l'opération. C’est une variable **quantitative discrète**.
- **Patient's year of operation (numérique, entier) :** Représente l'année au cours de laquelle le patient a subi l'opération. C'est une variable **quantitative discrète**.
- **Number of positive axillary nodes detected (numérique) :** Représente le nombre de ganglions axillaires positifs détectés chez le patient. C'est une variable **quantitative discrète**.
- **Survival status (numérique) :** Représente le statut de survie du patient après l'opération. Il s'agit d'une variable **qualitative binaire** avec deux catégories : 1 = le patient a survécu 5 ans ou plus et 2 = le patient est décédé dans les 5 ans.

3.Prétraitement

- Le jeu de données ne contient aucune valeur manquante.
- Seule la variable **Survival Status** est une variable catégorielle binaire (1 ou 2), celle-ci peut-être considérée comme une variable **supplémentaire** dans la suite de l’analyse de ce jeu de données.

4.Analyse univariée

4.1.Statistiques univariées

4.1.1.Variables quantitatives

4.1.1.1.Critères de position

Grâce à la commande “summary”, nous avons un aperçu globale des différentes distributions des variables de notre jeu de données en calculant les statistiques de base pour chaque variable (Voir les figures 1-2). Elle nous aide à détecter d'éventuelles anomalies dans leurs distributions et d’avoir quelque ordres de grandeurs de chaque attribut du jeu de données.

Age	Operation_year	Pos_axi_nodes
Min. :30.00	Min. :58.00	Min. : 0.000
1st Qu.:44.00	1st Qu.:60.00	1st Qu.: 0.000
Median :52.00	Median :63.00	Median : 1.000
Mean :52.46	Mean :62.85	Mean : 4.026
3rd Qu.:60.75	3rd Qu.:65.75	3rd Qu.: 4.000
Max. :83.00	Max. :69.00	Max. :52.000

Figure 1 : Résultats de la commande “summary” sur le jeu de données Haberman

```
$`1`
      Age        Year of operation NB positive axillary nodes
 Min.   :30.00   Min.   :58.00     Min.   : 0.000
 1st Qu.:43.00   1st Qu.:60.00     1st Qu.: 0.000
 Median :52.00   Median :63.00     Median : 0.000
 Mean   :52.02   Mean   :62.86     Mean   : 2.791
 3rd Qu.:60.00   3rd Qu.:66.00     3rd Qu.: 3.000
 Max.   :77.00   Max.   :69.00     Max.   :46.000

$`2`
      Age        Year of operation NB positive axillary nodes
 Min.   :34.00   Min.   :58.00     Min.   : 0.000
 1st Qu.:46.00   1st Qu.:59.00     1st Qu.: 1.000
 Median :53.00   Median :63.00     Median : 4.000
 Mean   :53.68   Mean   :62.83     Mean   : 7.457
 3rd Qu.:61.00   3rd Qu.:65.00     3rd Qu.:11.000
 Max.   :83.00   Max.   :69.00     Max.   :52.000
```

Figure 2 : Résultats de la commande “summary” sur le jeu de données Haberman divisé selon les valeurs de la variable “Survival Status”

4.1.1.2.Critères de dispersion

Variable	Etendue	Variance	Ecart-type	Conclusion
Age of patient at time of operation	53	116.7146	10.80345	La distribution des âges des patients au moment de l'opération montre une variation significative, avec une dispersion de 53 ans. Cela indique une diversité considérable dans les âges des patients dans l'échantillon.
Patient's year of operation	11	10.55863	3.249405	Les années d'opération des patients montrent une plage relativement étroite de 11 ans. Cela montre que les opérations ont été effectuées sur une période spécifique et limitée.
Number of positive axillary nodes detected	52	51.69112	7.189654	Le nombre de nœuds axillaires positifs détectés varie considérablement. Cela indique une grande disparité dans les résultats des tests de détection des nœuds axillaires positifs chez les patients.

4.1.1.3.Degré d'asymétrie

- **Age of patient at time of operation**

Le degré d'asymétrie de la distribution de l'âge des patients au moment de l'opération est de 0.1457859, ce qui indique une asymétrie positive. La distribution des âges est étirée vers les valeurs supérieures, ce qui signifie que la moyenne de l'âge est supérieure à la médiane.

- **Patient's year of operation**

Pour l'année de l'opération du patient, le degré d'asymétrie de la distribution est de 0.0779844. C'est une symétrie, avec une moyenne proche de 63 après arrondi. Cela indique que la distribution est relativement équilibrée autour de sa moyenne.

- **Number of positive axillary nodes detected**

Le degré d'asymétrie de la distribution est de 2.969176, indiquant une asymétrie positive. La distribution s'étale davantage vers les valeurs les plus élevées de la variable. Par conséquent, la moyenne du nombre de ganglions détectés est supérieure à la médiane.

4.1.1.4.Degré d'aplatissement

- **Age of patient at time of operation :** Le degré d'aplatissement de la distribution de l'âge des patients est de -0.6150152. Un aplatissement négatif indique une distribution platykurtique, avec des extrémités moins épaisses.
- **Patient's year of operation :** Pour l'année de l'opération du patient, le degré d'aplatissement de la distribution est de -1.132438. Un aplatissement négatif indique une distribution platykurtique, avec des extrémités moins épaisses.
- **Number of positive axillary nodes detected :** Le degré d'aplatissement de la distribution est de 11.42578. Un aplatissement positif indique une distribution leptokurtique avec des valeurs plus extrêmes et des extrémités fines.

4.1.2.Variables qualitatives

La seule variable qualitative est la variable "*Survival Status*", qui indique si un patient a survécu pendant 5 ans ou plus après l'opération (valeur 1) ou s'il est décédé dans les 5 ans suivant l'opération (valeur 2) (Voir la figure 5 en annexe).

4.1.2.1.Fréquence absolue

- Pour la valeur 1, indiquant la survie à long terme, la fréquence est de 225.
- Pour la valeur 2, indiquant la non-survie à long terme, la fréquence est de 81.

4.1.2.2.Fréquence relative

- Pour la valeur 1, la fréquence relative est de presque 73.53%.
- Pour la valeur 2, la fréquence relative est de presque 26.47%.

4.1.2.3.Interprétation

Il semble y avoir un plus grand nombre de patients ayant survécu pendant 5 ans ou plus par rapport à ceux qui sont décédés dans les 5 ans suivant l'opération. Cela montre que les données ne sont pas équilibrées. Cette différence pourrait être explorée pour comprendre les facteurs jouant sur la survie des patients.

4.1.3.Description des résultats obtenus

4.1.3.1.Histogrammes, densité de probabilité et fonction de répartition

4.1.3.1.1.Age of patient at time of operation

(Voir les figures 6-7 en annexe)

- Les patients âgés de 30 à 40 ont plus de chances de survie que les autres.
- Les patients âgés de 40 à 60 ont moins de perspectives de survie.
- Le plus grand nombre de décès se situe dans la tranche d'âge de 50 à 55 ans.

4.1.3.1.2. Patient's year of operation

(Voir les figures 8-9 en annexe)

- L'histogramme présente des informations sur la réussite et l'échec des opérations
- La plupart des opérations non réussies ont été réalisées entre 1958 et 1959, suivies de 1964 et 1965.
- La plupart des opérations réussies ont été réalisées entre 1959 et 1961.

4.1.3.1.3. Number of positive axillary nodes detected

(Voir les figures 10-11 en annexe)

- Les patients sans nœud axillaire positif ont des chances de survie très élevées par rapport aux patients qui en ont.
- Les patients avec un seul nœud axillaire positif ont également de bonnes chances de survie.
- La probabilité de survie au cancer du sein diminue avec l'augmentation du nombre de nœuds axillaires positifs détectés.

4.1.3.2.Boîtes à moustaches

Après avoir réalisé les boîtes à moustaches (voir les figures 12-13-14-15, en annexe), nous avons exploré en détail les caractéristiques de notre jeu de données. La variable qui a posé le plus de problèmes était le nombre de nœuds axillaires positifs détectés, tandis que les autres variables ne présentaient pas de valeurs aberrantes. Voici ce que nous avons découvert lors de cette analyse approfondie de cette variable :

- La présence de points en dehors de la moustache indique des valeurs aberrantes. Le nombre de valeurs aberrantes dans la catégorie "Survécu" est considérablement plus élevé que dans la catégorie "Décédé".
- Les valeurs de Q1 et de Q2 (médiane) pour la catégorie "Survécu" sont presque identiques. La médiane de la catégorie "Décédé" et le Q3 de la catégorie "Survécu" semblent alignés.
- La plupart des patients sans nœuds axillaires positifs ont survécu au cancer du sein, tandis que la majorité de ceux avec un nombre plus élevé de nœuds axillaires positifs sont décédés.
- Cependant, il existe des patients qui ont survécu avec un grand nombre de nœuds axillaires positifs et des patients sans nœuds axillaires positifs sont décédés.

5.Analyse bivariées

5.1.Corrélation

On remarque ici que la majorité des couples de variables sont faiblement corrélés (Voir la figure 16 en annexe).

L'âge et l'année d'opération présentent une corrélation positive faible (0,0895). Cela signifie que les patients plus âgés ont tendance à être opérés plus tard.

L'âge et le nombre de ganglions axillaires positifs présentent une corrélation négative faible (-0,0632), donc les patients plus âgés ont tendance à avoir un nombre moins élevé de ganglions axillaires positifs.

L'année d'opération et le nombre de ganglions axillaires positifs présentent une corrélation négative très faible (-0,0038). Cela signifie que les patients qui ont été opérés plus tard ont tendance à avoir un nombre moins élevé de ganglions axillaires positifs.

Le nombre de ganglions axillaires positifs et le statut de survie présentent une corrélation positive modérée (0,2868). Cela nous indique que les patients qui ont un nombre plus élevé de ganglions axillaires positifs ont tendance à avoir un moins bon pronostic de survie.

L'âge et le statut de survie présentent une corrélation positive faible (0,0680). Cela signifie que les patients plus âgés ont tendance à avoir un moins bon pronostic de survie.

	Age	Year of operation	NB positive axillary nodes
Age	1.00000000	0.089529446	-0.063176102
Year of operation	0.08952945	1.000000000	-0.003764474
NB positive axillary nodes	-0.06317610	-0.003764474	1.000000000

Figure 3 : Coefficients de corrélation des différents couples de variables du jeu de données Haberman

5.2.Graphique par paires

Puisque le jeu de données ne comprend que trois variables quantitatives, le graphique par paires (Voir la figure 17 en annexe) présente seulement 6 plots, chacun représentant une variable en fonction d'une autre.

En analysant ce dernier, on a pu observer cela:

- Il n'y a aucune séparation linéaire dans aucun des graphiques, ce qui veut dire qu'il n'existe pas de relation linéaire entre les couples de variables.Cela peut indiquer une complexité dans la relation ou une absence de séparation exacte entre les données.
- Dans les plots (1, 2) et (2, 1), qui représentent l'âge du patient en fonction de l'année de l'opération et l'année de l'opération en fonction de l'âge du patient respectivement, il y a un chevauchement remarkable qui rend difficile la division.
- Dans les plots (1, 3) et (3, 1), qui représente l'âge du patient en fonction du nombre de nœuds axillaires positifs détectés et le nombre de nœuds axillaires positifs détectés en fonction de l'âge du patient respectivement, il y a un chevauchement de points, mais il existe des patterns distinguables qui nous permettent d'analyser et de déduire que le nombre de nœuds axillaires positifs détectés et l'âge du patient sont les caractéristiques les plus utiles pour identifier le statut de survie d'un patient.
- Dans les plots (2, 3) et (3, 2), qui représentent le nombre de nœuds axillaires positifs détectés en fonction de l'année de l'opération et l'année de l'opération en fonction du nombre de nœuds axillaires positifs détectés respectivement, il y a un chevauchement remarkable qui rend difficile la division.

6.Analyse multivariées

6.1.PCA

Dans le graphe des variables (Voir la figure 18 en annexe), on remarque que la variable “Age” est fortement corrélée avec le premier axe, cela signifie que la variable contribue de manière significative à la définition du premier axe de l'ACP.

Les 2 axes ensemble représentent 70.27% de la variance totale des données. Autrement dit, environ 70.27% de la variation observée dans les données peut être expliquée par ces deux axes principaux.

6.2.Classification

Afin de mieux comprendre notre base de données, nous pouvons utiliser des méthodes de groupement. Cela permet de regrouper les individus qui ont une certaine similitude afin de mieux comprendre et prédire les données futures. Une des premières méthodes de regroupement est la classification ascendante hiérarchique.

6.2.1.Classification ascendante hiérarchique (CAH)

En raison d’un nombre d’individu relativement bas (306), nous avons pu faire une CAH, et l'interprétation de celle-ci a été raisonnable. Nous avons ici utilisé pour métrique la distance euclidienne.

Sur le dendrogramme (Voir les figures 19-20 en annexe), nous pouvons observer le regroupement des individus en plusieurs classes (4,8 ou 32). Il y a une multitude de regroupements qui ne seront pas l'objet de notre étude, nous nous contenterons d'étudier le regroupement sous 4 classes pour simplifier. Nous observons 4 classes fortement distinctes encadrées sur la figure en vert.

Le dendrogramme coupé (Voir la figure 21 en annexe) est aussi une manière d’observer le nombre de classes à considérer : Il suffit de prendre en compte les sauts d’inerties, c’est a dire observer à la l'ordonnée si il y a une différence élevée de l’inertie. On aurait pu prendre plus de cours mais nous nous sommes tenus à 4 classes.

Nous stockerons ces 4 classes dans 4 variables différentes (de gauche à droite sur le dendrogramme) (Voir les tableaux figures 22-23 en annexe).

Nous avons utilisé la fonction catdes() sur la variable de classe nous a permis de faire des études statistiques au sein de chaque classe pour appuyer nos propos, ainsi que de summary() pour completer les donnees.

De plus observons les valeurs de eta2, dont une valeur haute montre une forte corrélation entre les variables quantitatives (ici X1.1, X1, X64 et X30) et la variable qualitative qui ici est la classe. Nous remarquons dans la figure 4, X1.1 est fortement corrélée avec la classe, puis le nombre de nœuds de manière remarquable. De manière complémentaire, la P-value montre la probabilité de l'hypothèse nulle (qu'elle ne soit pas impliquée dans cette classification).

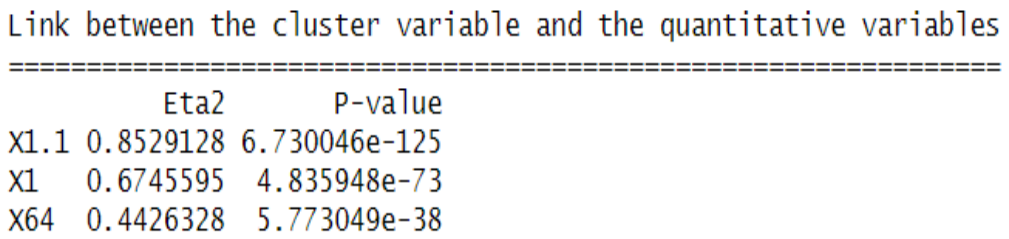

Figure 4 : Coefficients de corrélation entre la variable de classe et des variables quantitatives

6.2.2.Classification non hiérarchique (K-moyenne)

Dans le cas d'étude d'une grande quantité de données, il est toujours possible d'utiliser une classification non hiérarchique afin d'obtenir des résultats dans un temps raisonnable en omettant certaines informations. Ici on prend la peine tout de même d'utiliser la technique de classification afin de comparer ces résultats avec le CAH.

Tout d'abord on va essayer de déterminer le nombre de clusters idéal, en supposant que le nombre de clusters maximum est 8 (Voir la figure 24 en annexe).

Nous observons un nombre de clusters idéal de 2. Cependant, nous prenons le choix de garder le nombre de clusters a k=4 afin de pouvoir comparer avec le CAH. Puis nous allons afficher les clusters et quels points sont associés à quels clusters, le graphique montre que les points sont très mêlées entre eux, il est donc assez difficile de les distinguer de manière impartiale (Voir la figure 25 en annexe).

Ensuite, de même que pour le CAH, nous allons créer 4 tableaux de "classes" liés à chaque clusters, et effectuer les études statistiques (Voir les tableaux figures 26-27-28 dans l'annexe).

7.Conclusion

Suite à ces deux analyses de classifications des individus atteints du cancer du sein, on remarque que les deux sont tout à fait possibles. Cependant, la CAH semble apporter plus de subtilités ce qui est cohérent car il s'agit de la classification prenant en compte le plus de facteurs.

La classification avec K-means avec k = 4 regroupe tous les patients morts avant 5 ans sans différenciation, alors qu' avec CAH, on fait la différence entre les morts inattendues (individus avec peu de nœuds, qui ont eu une opération) et ceux qui ont beaucoup de noeuds. Ce dernier met en évidence également qu'il y a grande chance de morts s'il y a beaucoup de nœuds, mais qu'il y a quand même une chance faible de mourir prématurément du cancer.

Annexe

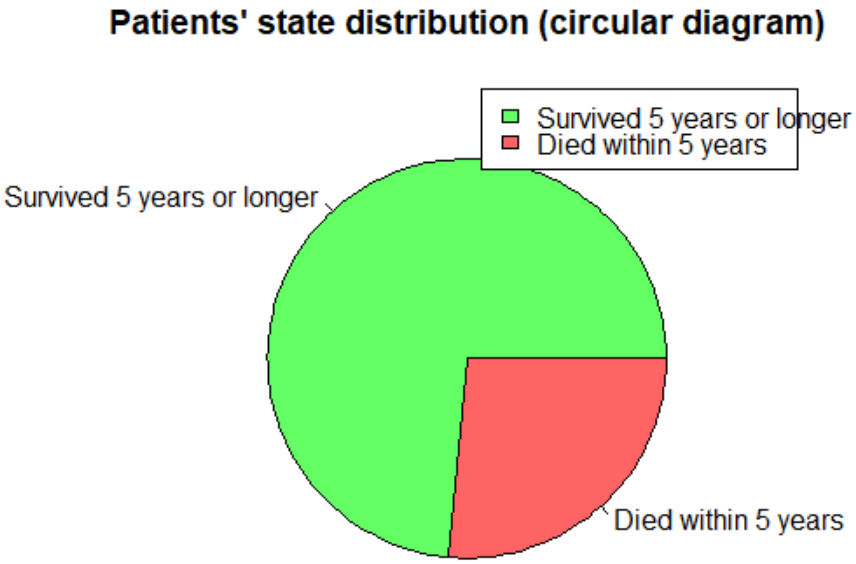

Figure 5 : Diagramme circulaire représentant l'état des patients

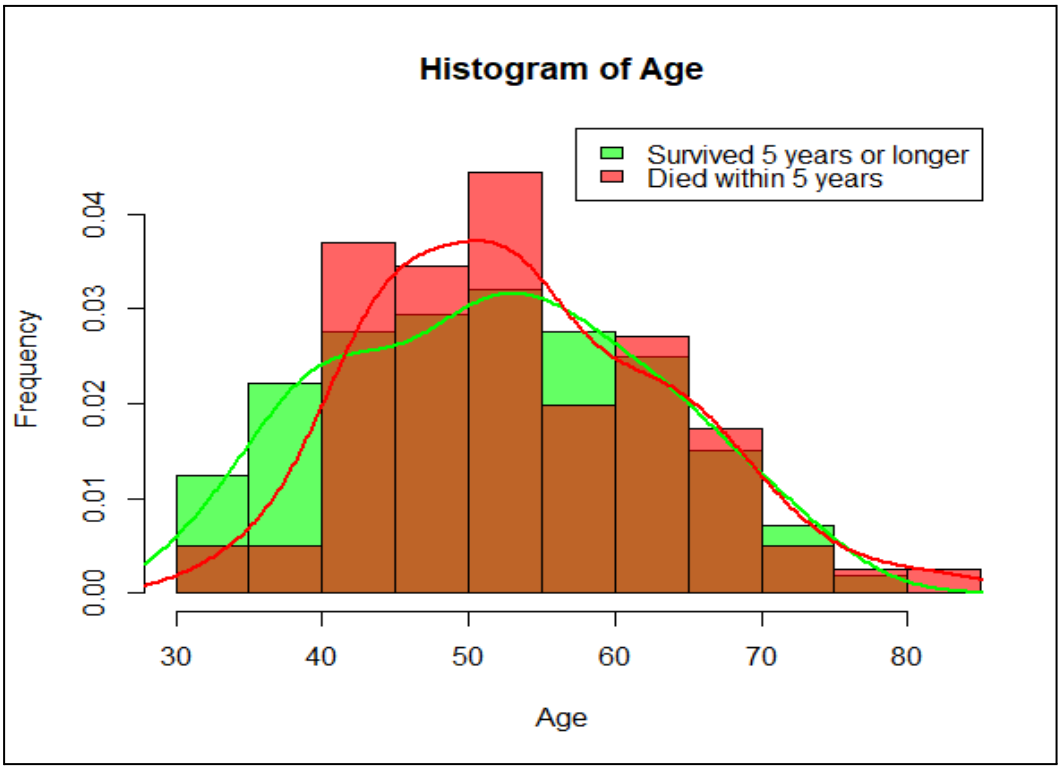

Figure 6 : Histogramme et densité de probabilité d'âge

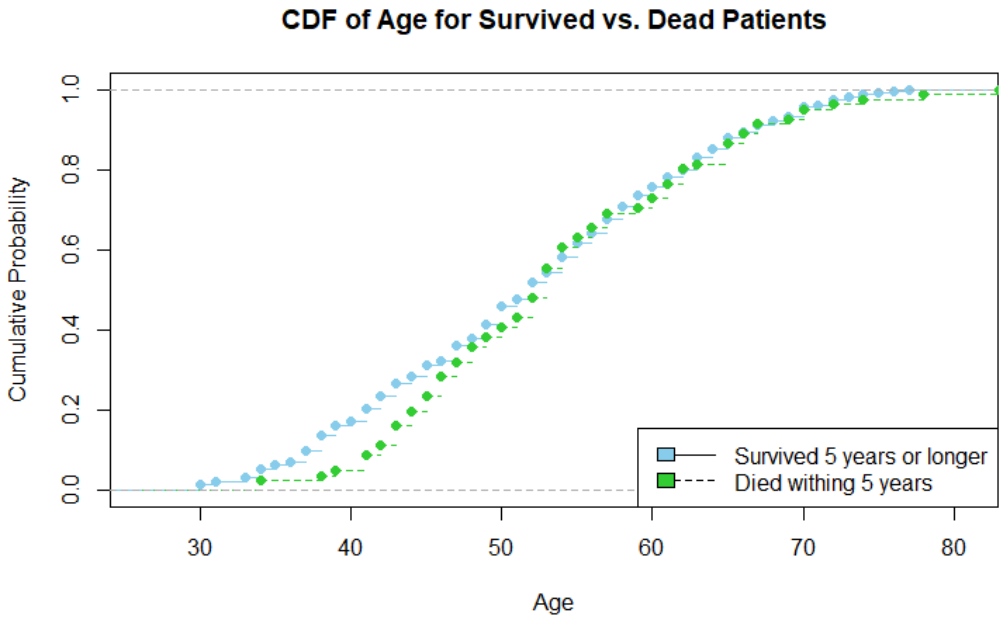

Figure 7 : Fonction de répartition d'âge

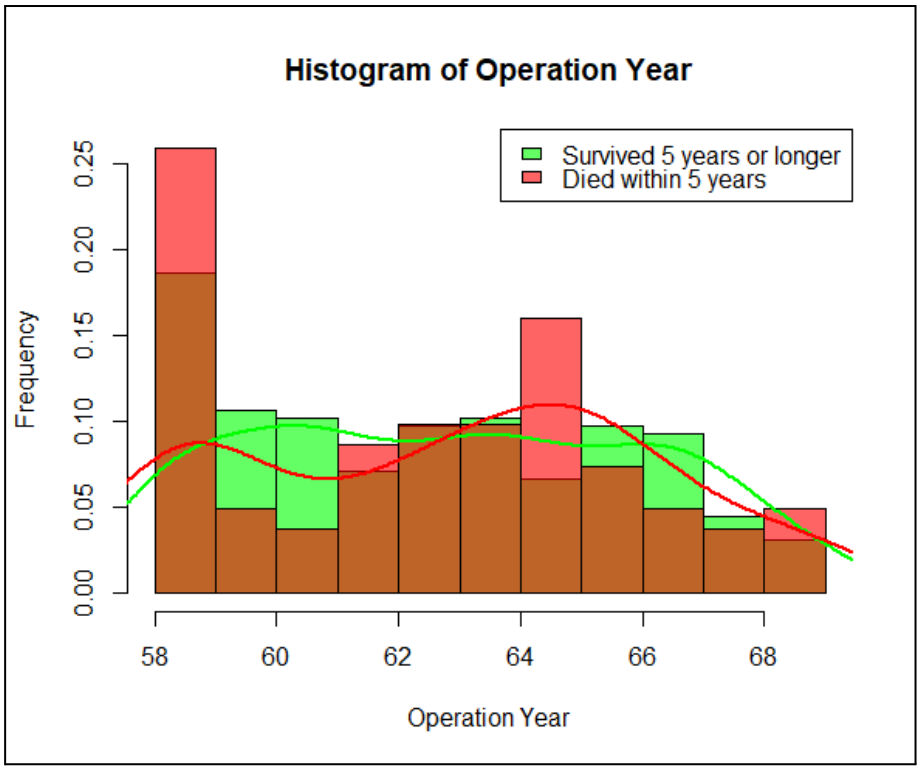

Figure 8 : Histogramme et densité de probabilité des années d'opération

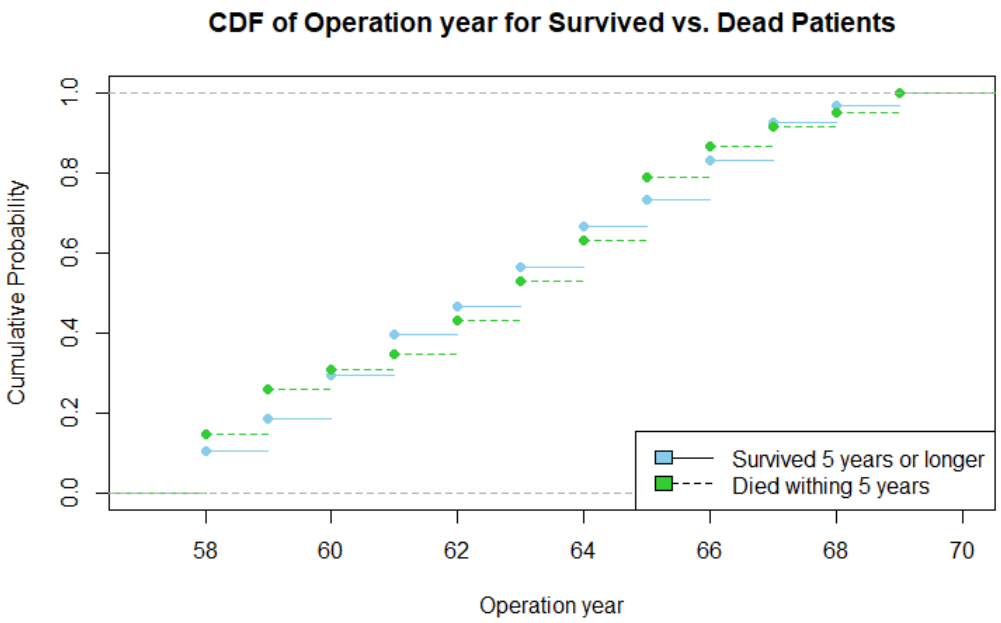

Figure 9 : Fonction de répartition des années d'opération

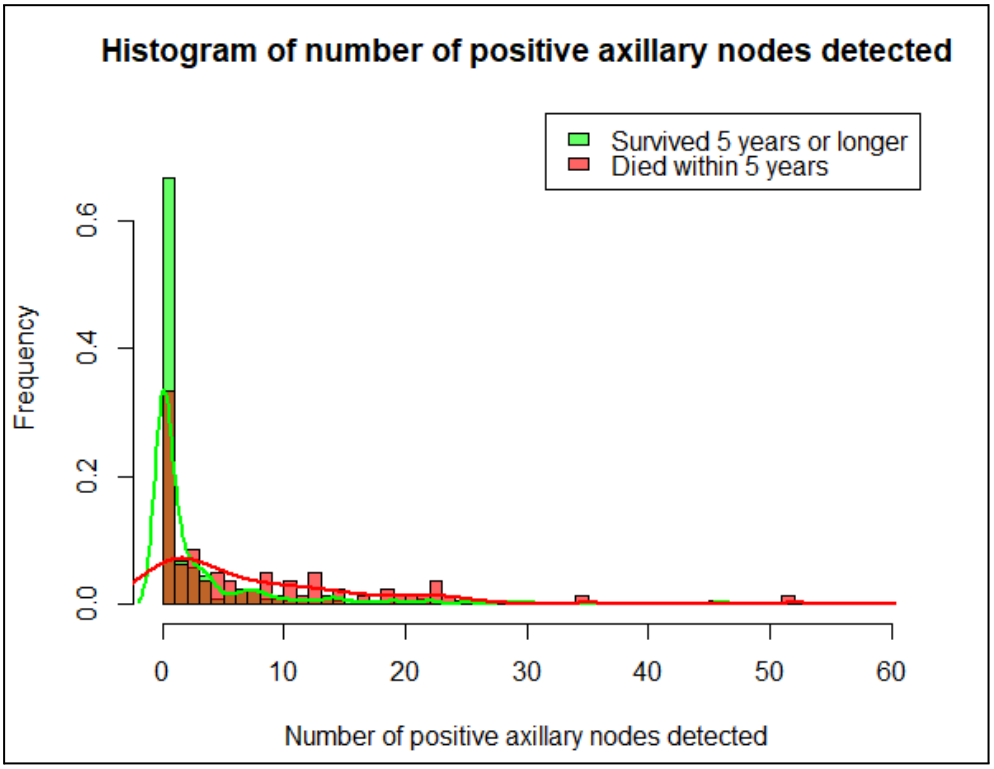

Figure 10 : Histogramme et densité de probabilité des nombre de noeuds positifs détectés

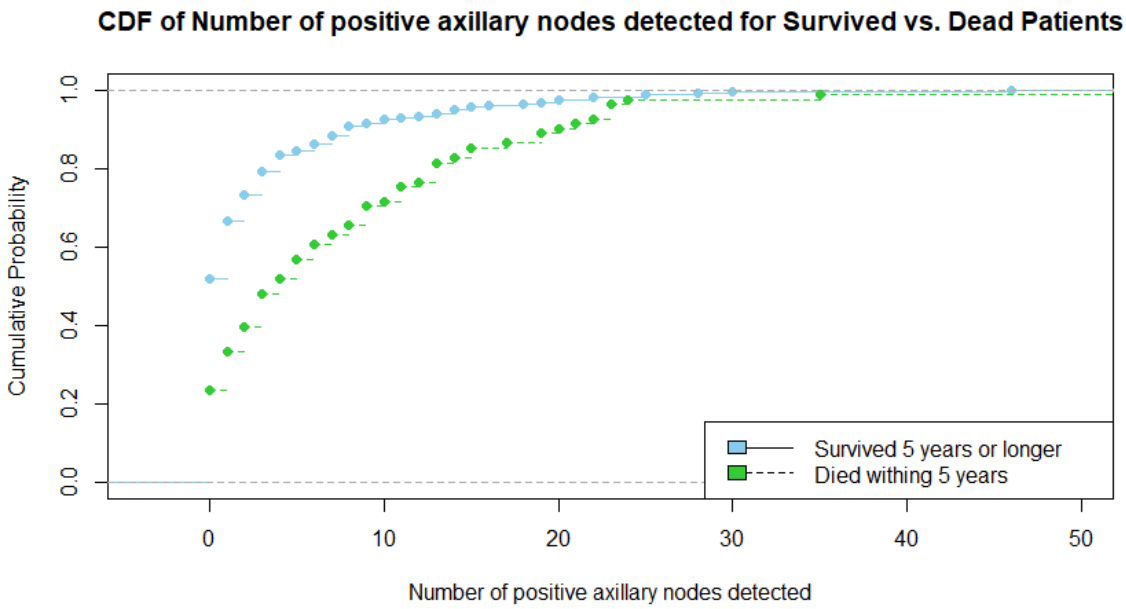

Figure 11 : Fonction de répartition des nombre de noeuds positifs détectés

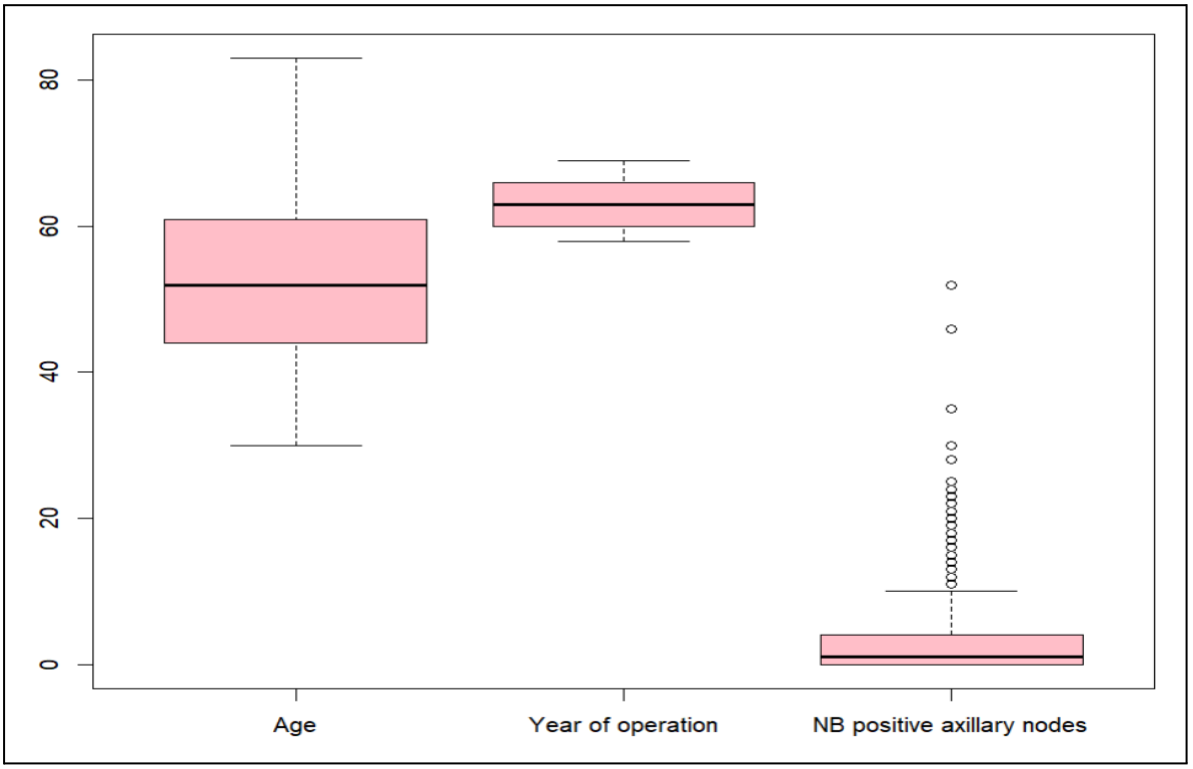

Figure 12 : Boîtes à moustaches des variables quantitatives du jeu de données "Haberman"

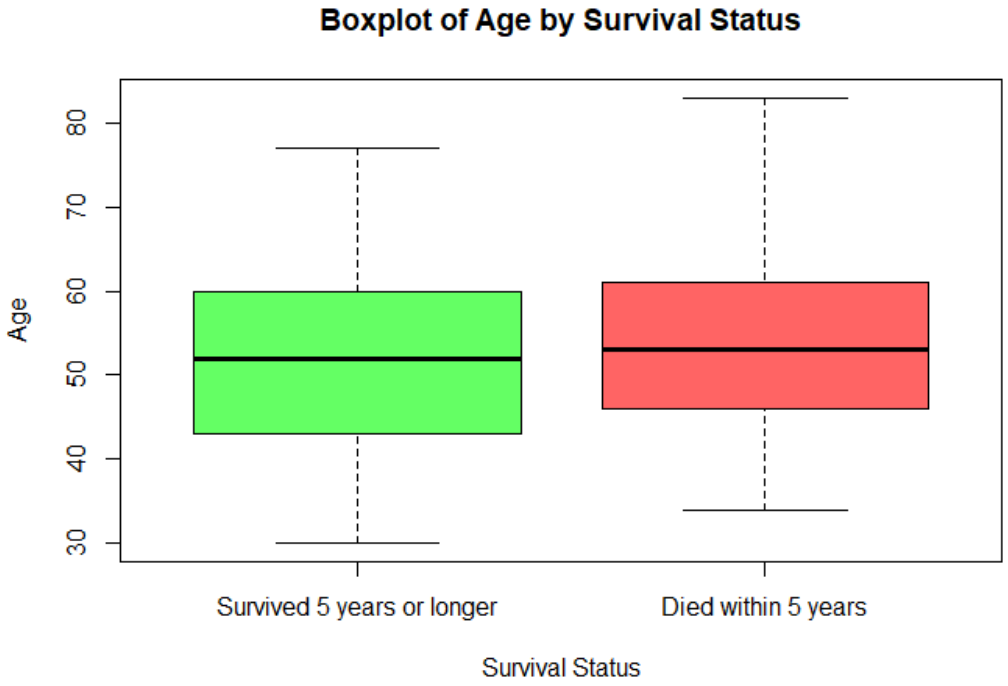

Figure 13 : Boîte à moustaches d'âge

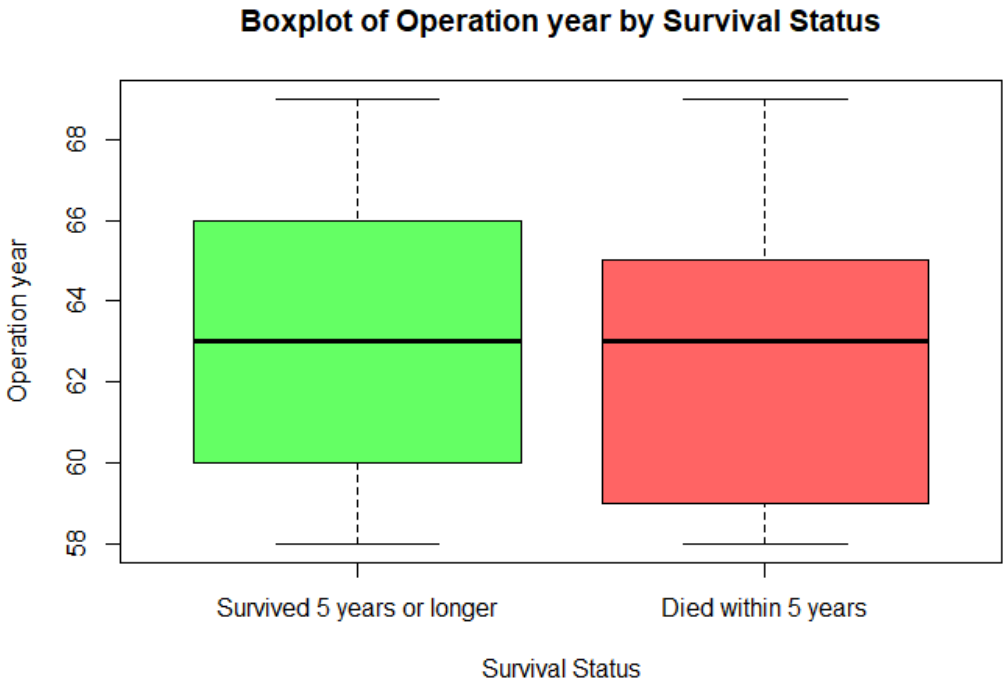

Figure 14 : Boîte à moustaches d'année d'opération

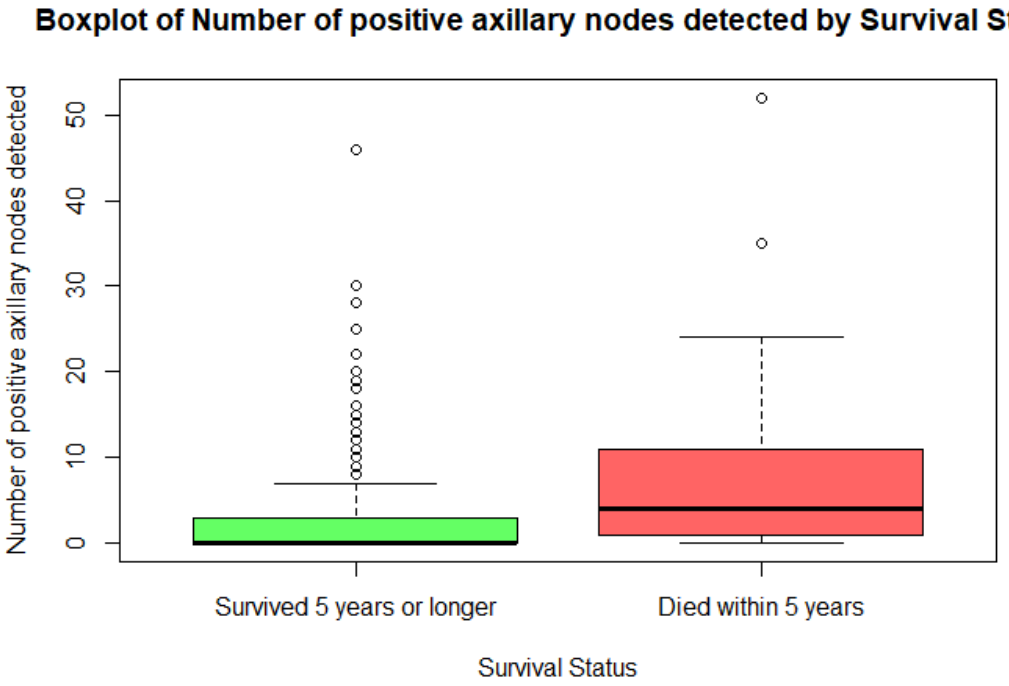

Figure 15 : Boîte à moustaches des nombre de noeuds positifs détectés

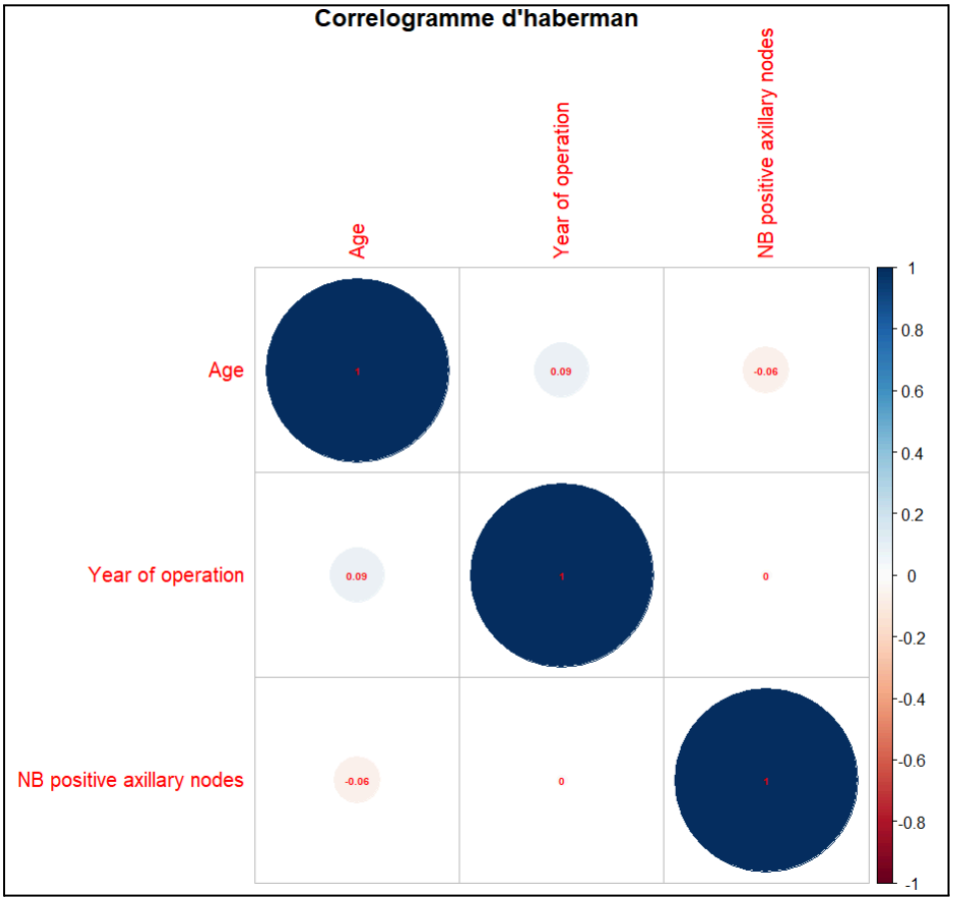

Figure 16 : Corrélogramme du jeu de données “Haberman”

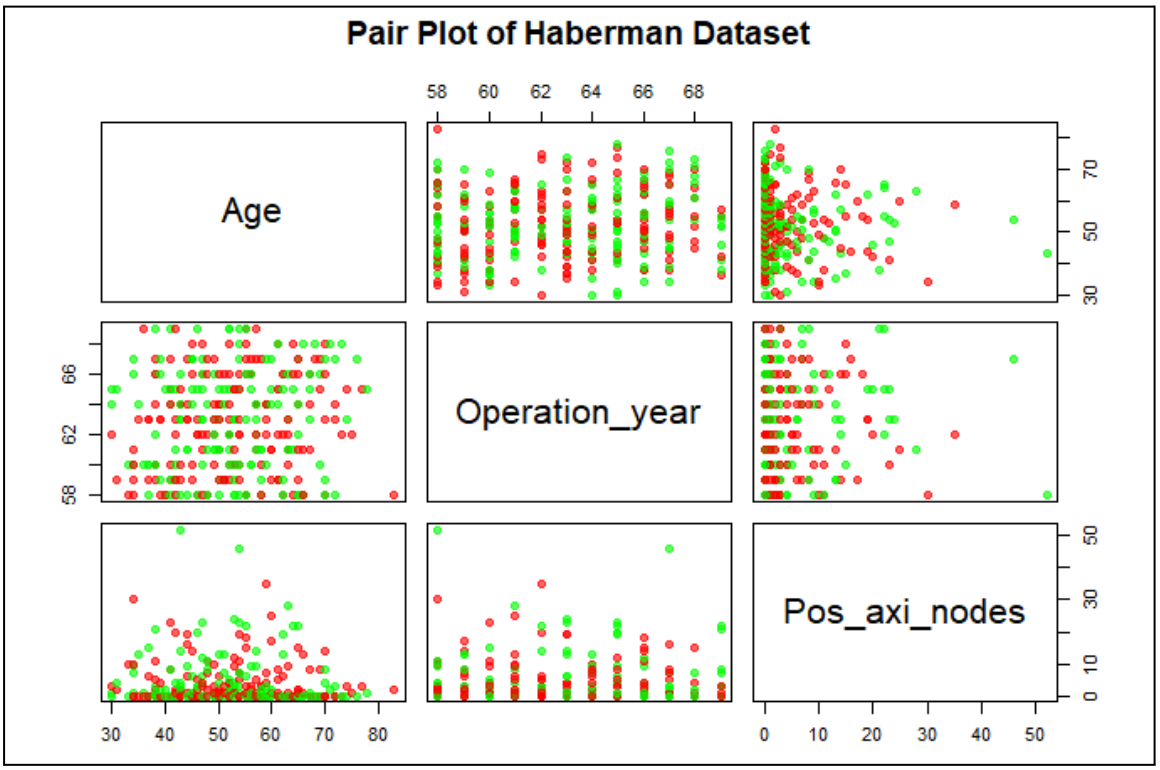

Figure 17 : Graphique par paires

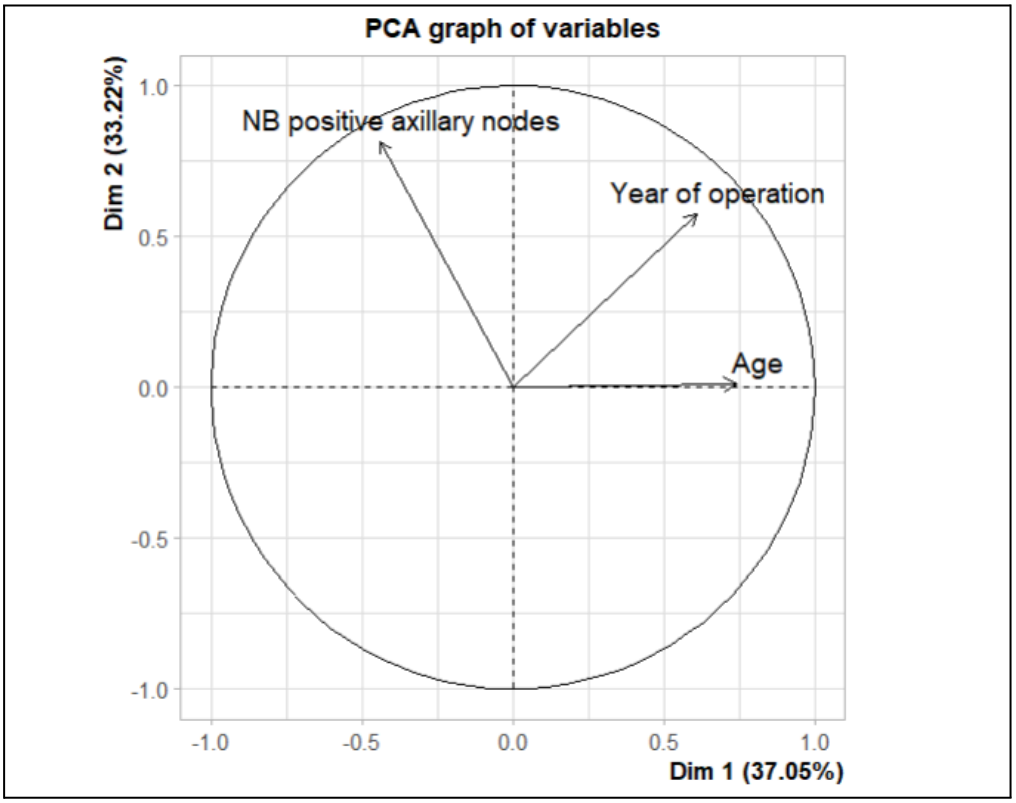

Figure 18 : Graphe des variables du PCA

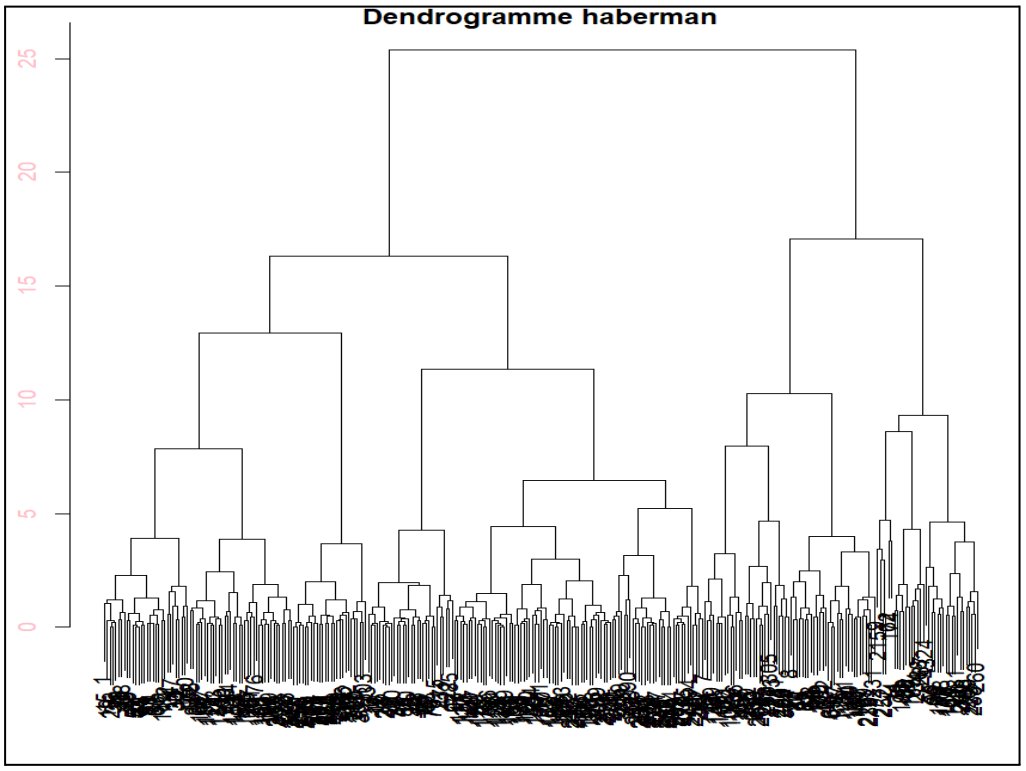

Figure 19 : Dendrogramme du CAH

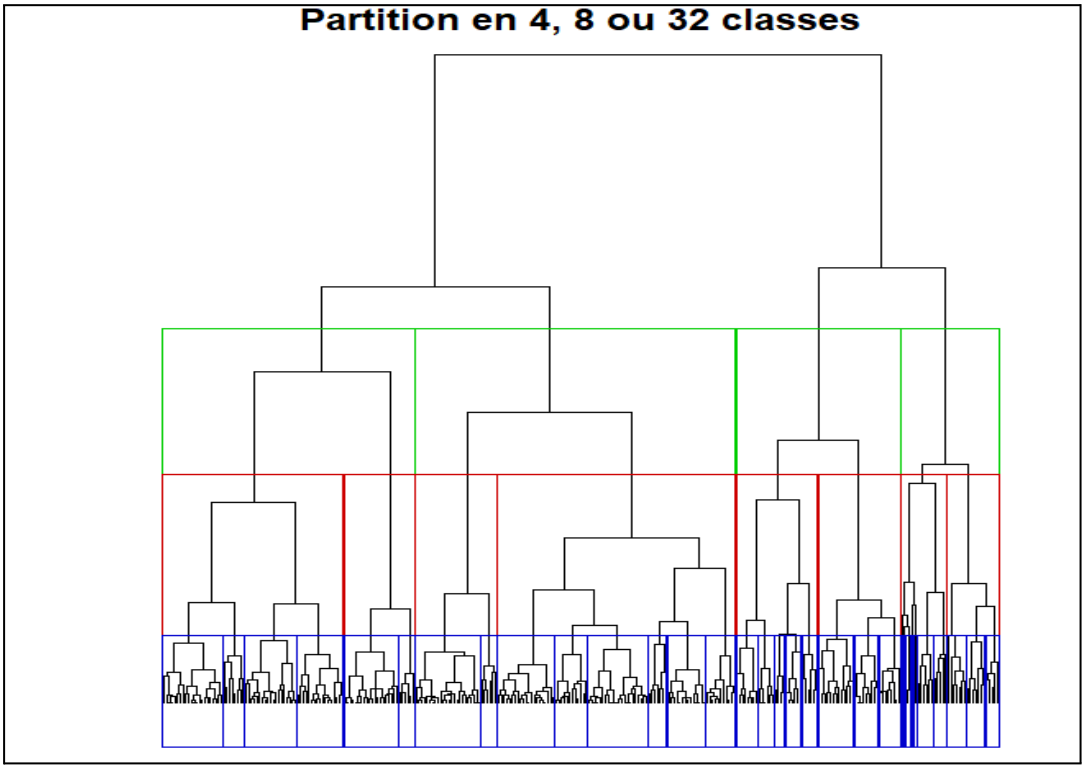

Figure 20 : Dendrogramme du CAH avec encadrement des classes

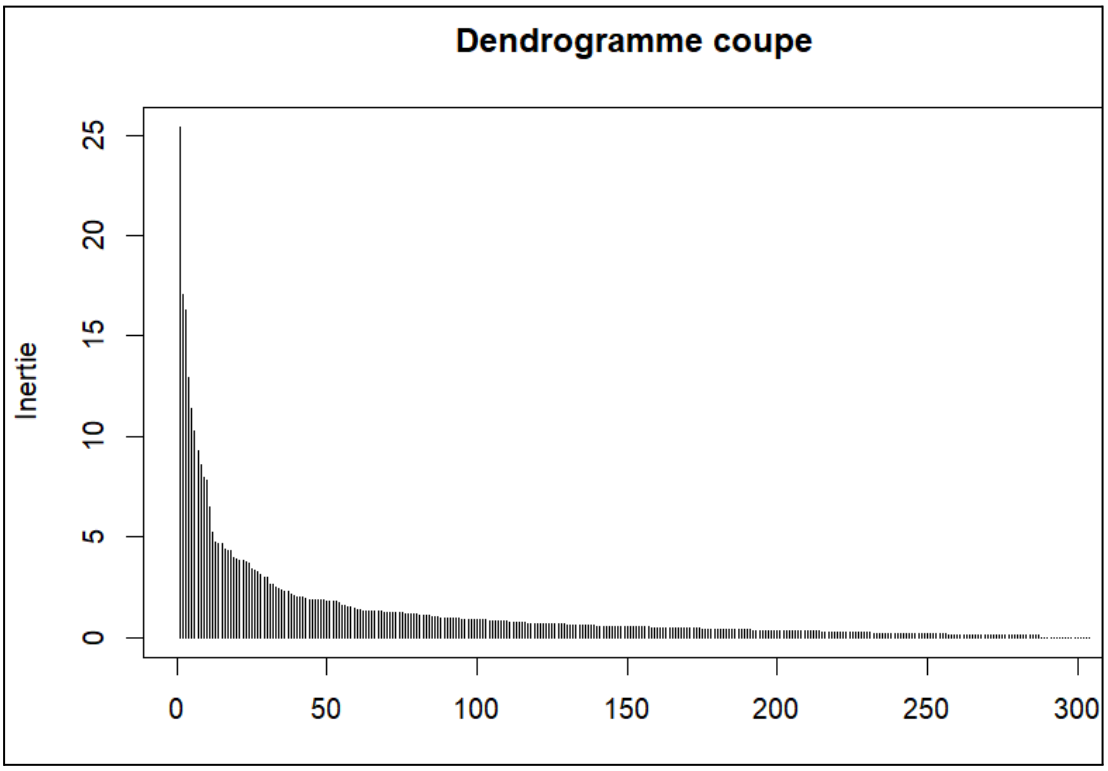

Figure 21 : Dendrogramme coupé du CAH

Classes	Similitudes
Classe 1 (92 individus)	100% d'individus ayant tous survécu au moins plus de 5 ans. Individus ayant globalement très peu de nœuds. `(Cluster $`1`, X1: Mean in category 1.152174)` Individus ayant effectué l'opération entre 1962 et 1969. `(Cluster $`1`, X64: Mean in category 65.673913)` Pas de lien spécial pour l'âge. `(mean = 52.75)`
Classe 2 (117 individus)	100% d'individus ayant tous survécu au moins plus de 5 ans. Individus ayant globalement très peu de nœuds, mais plus que la classe 1. `(Cluster $`2`, X1: Mean in category 1.803419)` Individus ayant effectué l'opération entre 1958 et 1964. `(Cluster $`2`, X64: Mean in category 60.461538)` Pas de lien spécial pour l'âge. `(mean = 51.82)`
Classe 3 (60 individus)	100% d'individus n'ayant pas survécu plus de 5 ans. Individus ayant globalement très peu de nœuds, mais plus que la classe 1 et 2.

	(mean = 52.75) Dates d'à peu près toutes les périodes. (mean = 63.64) Pas de lien spécial pour l'âge. (mean = 53.72)
Classe 4 (36 individus)	La moitié des individus ont survécu plus de 5 ans, l'autre moitié non. (Cluster $`4`, X1.1: Mean in category 1.583333) Individus ayant globalement un grand nombre de nœuds. (Cluster $`4`, X1: Mean in category 20.055556) Dates d'à peu près toutes les périodes. (mean = 62.7) Pas de lien spécial pour l'âge. (mean = 52.31)

Figure 22 : Tableau des similitudes CAH

Classes	**Conclusions**
Classe 1	Patients ayant subi une opération vers 1965, ayant eu très peu de nœuds détectés positifs au cancer. **Ce sont des patients qui ont probablement bénéficié d'une survie prolongée suite à une opération réussie.**
Classe 2	Patients ayant subi une opération plus tôt, vers 1960. Ils ont eu également très peu de nœuds détectés positifs au cancer. **Ce sont des patients qui ont probablement bénéficié d'une survie prolongée suite à l'opération, mais dont il faut potentiellement surveiller** certains en raison de la date ancienne et d'une présence un peu plus marquée de nœuds chez certains patients.
Classe 3	Patients de toutes les dates, ayant eu très peu de nœuds détectés positifs au cancer également. Cependant, ce sont des patients dont l'opération s'est peut être mal déroulée ou bien n'a pas pu empêcher efficacement la propagation du cancer. **Ce sont des cas qui auraient pu difficilement être prédits en avance.** *Remarque : Classe1/(Classe1+classe3) = 92/(92+60) = 60,5%* *Donc 60,5% des patients ayant peu de nœuds survivent, et 39,5% ont des complications malgré cela.*
Classe 4	Patients de toutes les dates, ayant eu un grand nombre de nœuds détectés positifs. **Leur pronostic vital était assez mauvais**. Certains patients, par chance, ont eu une opération qui a prolongé leur survie, mais la majorité (58,3% des patients de classe 4) périssent avant 5 ans suivant l'opération.

Figure 23 : Tableau des conclusions CAH

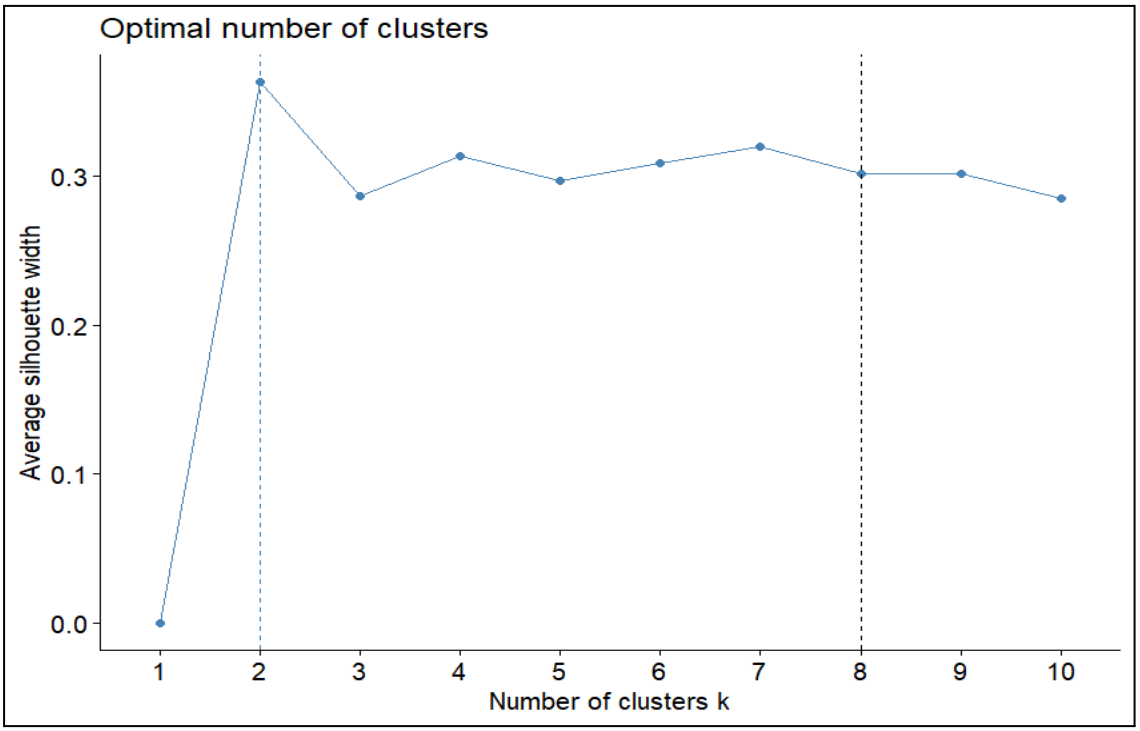

Figure 24 : Nombre optimal de clusters

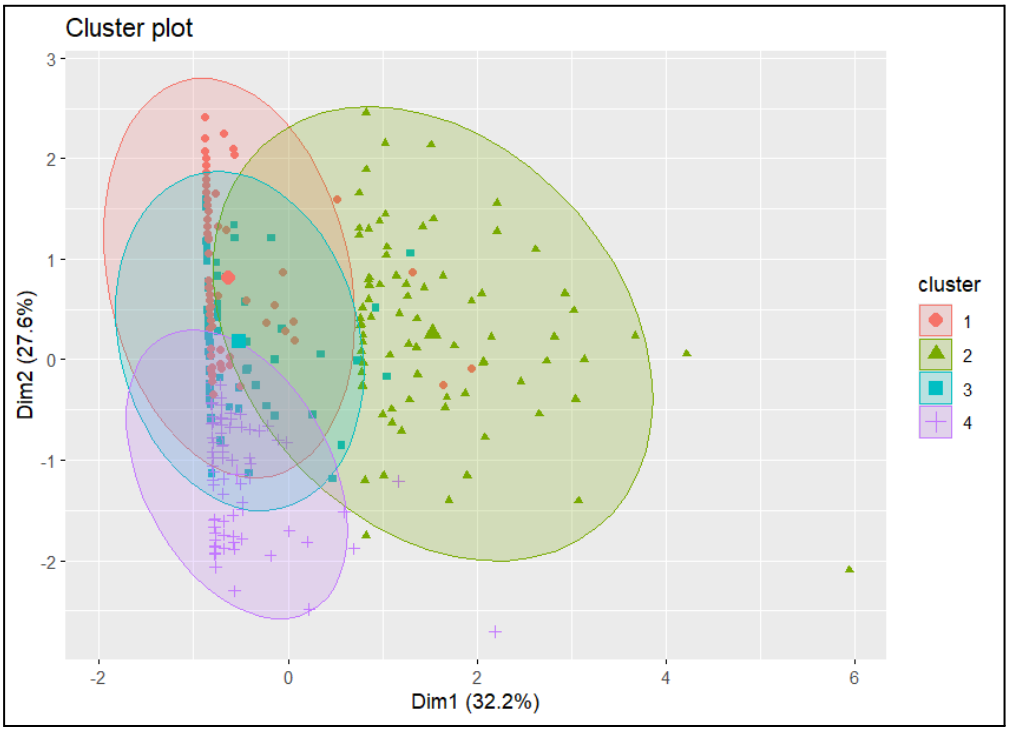

Figure 25 : CAH tableaux de classes

Analyse: Quels sont les liens intra-classes?

Rappel: Étude menée entre 1958 et 1970

Classes	Similitudes
Classe 1 (76 individus)	100% d'individus ayant tous survécu au moins plus de 5 ans. Individus ayant globalement très peu de nœuds. `(Cluster $`1`, X1: Mean in category 2.144737)` Individus ayant effectué l'opération entre 1958 et 1968. `(mean = 63.4)` Âge des individus élevés entre 53 et 77. `(Cluster $`30`, X30: Mean in category 64.065789)`
Classe 2 (82 individus)	100% d'individus n'ayant pas survécu plus de 5 ans. Individus ayant globalement pas mal de nœuds, mais plus que la classe 1 et 2. `(Cluster $`1`, X1: Mean in category 7,926829)(et variance de 10.0107022)` Individus ayant effectué l'opération entre 1958 et 1969. `(mean = 62.88)` Pas de lien spécial pour l'âge (38-83). `(mean = 64.07)`
Classe 3 (64 individus)	100% d'individus ayant tous survécu au moins plus de 5 ans. Individus ayant globalement très peu de nœuds, mais plus que la classe 1. `(mean = 3.266)` Individus ayant effectué l'opération entre 1963 et 1969. `(Cluster $`3`, X64: Mean in category 65.90625)` Individus assez jeunes (30-61). `(Cluster $`3`, X30: Mean in category 47.34375)`
Classe 4 (83 individus)	100% d'individus ayant tous survécu au moins plus de 5 ans. Individus ayant globalement un grand nombre de nœuds. `(Cluster $`4`, X1: Mean in category 2.518072)` Individus ayant effectué l'opération entre 1958 et 1963. `(Cluster $`4`, X64: Mean in category 60.108434)` Pas de lien spécial pour l'âge (30-58). `(Cluster $`4`, X30: Mean in category 44.831325)`

Figure 26 : Tableau des similitudes du K-means

Remarque : La mention "mean =" signifie que summary() a été utilisé, sinon nous avons utilisé catdes().

```
Link between the cluster variable and the quantitative variables
================================================================
          Eta2       P-value
X1.1 0.9833950 1.925332e-267
X30  0.4805504  1.492808e-42
X64  0.3830567  2.334188e-31
X1   0.1105839  1.036452e-07

Description of each cluster by quantitative variables
======================================================
$`1`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
X30  10.801302        64.065789    52.531148       5.480433 10.7263967 3.393570e-27
X1   -2.643087         2.144737     4.036066       5.355037  7.1875583 8.215388e-03
X1.1 -6.040103         1.000000     1.265574       0.000000  0.4416382 1.540158e-09

$`2`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
X1.1 17.290231         1.987805     1.265574      0.1097561  0.4416382 5.572843e-67
X1    5.723277         7.926829     4.036066     10.0107022  7.1875583 1.044887e-08

$`3`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
X64   8.454907         65.90625    62.849180       1.747487  3.2487386 2.793104e-17
X30  -4.345242         47.34375    52.531148       7.155022 10.7263967 1.391224e-05
X1.1 -5.403022          1.00000     1.265574       0.000000  0.4416382 6.552732e-08

$`4`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
X1   -2.251582         2.518072     4.036066       4.709287  7.1875583 2.434868e-02
X1.1 -6.410883         1.000000     1.265574       0.000000  0.4416382 1.446792e-10
X30  -7.652903        44.831325    52.531148       7.280638 10.7263967 1.964926e-14
X64  -8.994012        60.108434    62.849180       1.605803  3.2487386 2.383664e-19
```

Figure 27 : Résultat de catdes()

Interprétation: Conclusions possibles pour chaque classe?

Classe	Conclusions
Classe 1	Patients ayant subi une opération vers 1963, ayant eu très peu de nœuds détectés positifs au cancer. Leur âge est assez grand (53-77). **Ce sont des vieux patients qui ont probablement bénéficié d'une survie prolongée suite à une opération réussie.**
Classe 2	Patients de toutes les dates, ayant eu un plutôt grand nombre de nœuds détectés positifs. **Ce sont tous les patients n'ayant pas survécu très longtemps apres leur opération, qu'ils aient eu beaucoup de nœuds ou non.**
Classe 3	Patients très jeunes, la plupart ayant eu très peu de nœuds détectés positifs au cancer également. Une minorité ont beaucoup de nœuds mais leur opération a été

	très bénéfique tout de même. Ils sont opérés plutôt tard pendant la période, aux environs de 1965. **Les jeunes patients aux environs de 1965 ont probablement bénéficié d'une survie prolongée suite à une opération réussie.**
Classe 4	Patients très jeunes, la plupart ayant eu très peu de nœuds détectés positifs au cancer également. Une minorité ont beaucoup de nœuds mais leur opération a été très bénéfique tout de même. Ils sont opérés plutôt tard pendant la période, aux environs de 1960.. **Les jeunes patients aux environs de 1960 ont probablement bénéficié d'une survie prolongée suite à une opération réussie.**

Figure 28 : Tableau des conclusions CAH