Université Paris Cité

Rapport Final
Projet de Traitement numérique de données

Année : 2024-2025

Groupe N°3 :

GOURMELEN Thomas, HAMDAN Nadil, PONNOUSSAMY Valentin

Enseignant : Nicoleta Rogovschi

Sommaire :

1. Objet de l'analyse

Le but de l'analyse de ces données est d'étudier le jeux de donnés qui nous à été fourni. Ce jeux de donnés contient les données d'une étude, réalisée de 1958 et 1970 à l'université de Chicago, portant sur la survie des patients ayant subi une intervention chirurgicale pour un cancer du sein. L'objectif de ce projet est de mettre en relation les différentes données en notre possession sur cette étude.

2. Description des données

Le jeux de donnée contient 306 observations de patients. Chaque observation contient 4 variables.

1. Âge du patient au moment de l'opération (numérique).
2. Année de l'opération (exprimée en année -1900, numérique).
3. Nombre de ganglions axillaires positifs (c'est-à-dire dans lesquels des cellules cancéreuses ont été détectées). (numérique).
4. Statut de survie :
    1. 1 = Le patient survie 5 ans ou plus.
    2. 2 = le patient décède dans les 5 ans.

2.1. Vérification des données :

2.1.1. Vérification du manque de donnée :

En utilisant la fonction : `colSums(is.na(data))`

Nous avons ce résultat :

```
                                Age du patient
                                             0
                     Annee de l'operation -1900
                                             0
Nombre de ganglions axillaires positifs detectes
                                             0
                                Statut de survie
                                             0
```

Cela nous indique que aucune donnée n'est manquante.

2.1.2. Vérification du type des variables :

En utilisant la fonction : `str(data)`

Nous avons ce résultat :

```
'data.frame':   306 obs. of  4 variables:
 $ Age du patient                              : int  30 30 30 31 31 33 33 34 34 34 ...
 $ Annee de l'operation -1900                  : int  64 62 65 59 65 58 60 59 66 58 ...
 $ Nombre de ganglions axillaires positifs detectes: int  1 3 0 2 4 10 0 0 9 30 ...
 $ Statut de survie                            : int  1 1 1 1 1 1 1 2 2 1 ...
```

Nous savons maintenant que toutes les variables sont du bon type.

2.1.3. Vérification des valeurs aberrantes :

En utilisant la fonction : `summary(data)`

Nous avons ce résultat :

```
Age du patient  Annee de l'operation -1900
Min.   :30.00   Min.   :58.00
1st Qu.:44.00   1st Qu.:60.00
Median :52.00   Median :63.00
Mean   :52.46   Mean   :62.85
3rd Qu.:60.75   3rd Qu.:65.75
Max.   :83.00   Max.   :69.00
Nombre de ganglions axillaires positifs detectes Statut de survie
Min.   : 0.000                                   Min.   :1.000
1st Qu.: 0.000                                   1st Qu.:1.000
Median : 1.000                                   Median :1.000
Mean   : 4.026                                   Mean   :1.265
3rd Qu.: 4.000                                   3rd Qu.:2.000
Max.   :52.000                                   Max.   :2.000
```

Les valeurs aberrantes pouvant être repérées en effectuant un résumé statistique du jeu de donnée. Nous observons ici qu'aucune aberration n'est présente.

3. Analyse

3.1. Analyse univariée

L'analyse univariée permet d'étudier la distribution de chaque variable individuellement. Pour une analyse univariée nous pouvons dans un premier temps faire un résumé statistique. Cette analyse à été réalisée dans la partie 2.3.

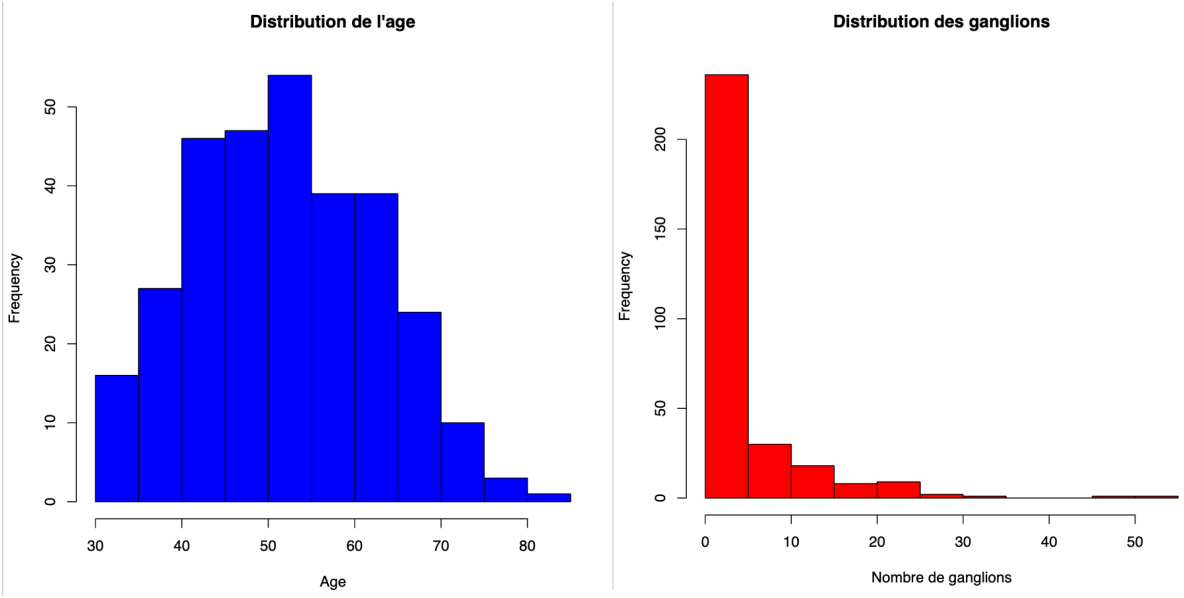
	Min	Max	Médiane	Moyenne
Age	30 ans	83 ans	52 ans	52,46 ans
Année	1958	1969	1963	1962
Nbr ganglions	0	52	1	4
Statut de survie	X	X	1	1,265

Cette analyse univariée permet de constater, grâce à la médiane, que plus de 50% des patients survive au-delà de 5 ans après l'opération.

Dans un second temps nous avons visualisé la distribution des valeurs en utilisant la fonction :

```
hist(data$Age, main = "Distribution de l'age", xlab = "Age", col = "blue")
hist(data$`Nombre de ganglions axillaires positifs`,
     main = "Distribution des ganglions",
     xlab = "Nombre de ganglions", col = "red")
```

Cette fonction nous permet de visualiser les valeurs sous forme d'histogramme :

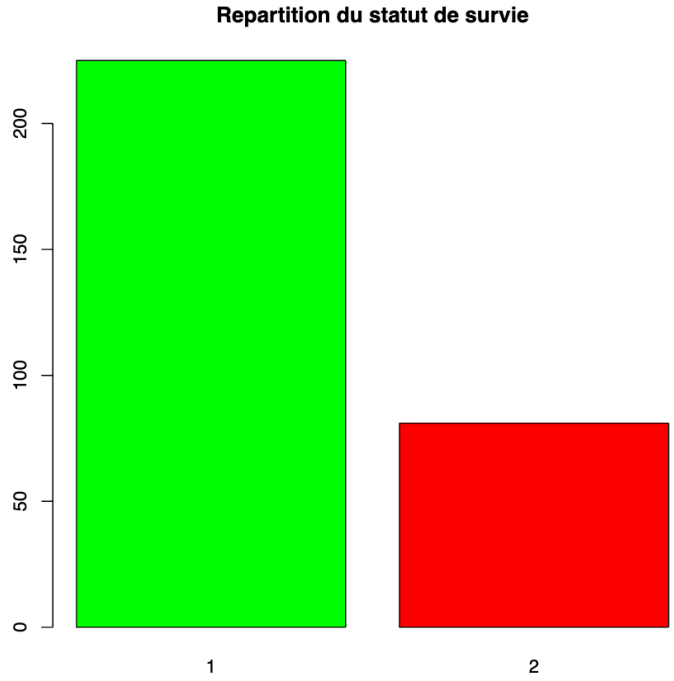


Grâce à la visualisation de la distribution de l'âge nous pouvons remarquer que la fourchette d'âge la plus représenté dans cette étude se situe entre 50 et 55 ans. Il y'a plus de 50 patients dans ce cas. Nous pouvons également remarqué que dans cette étude la majorité des patients ont entre 0 et 5 ganglions axillaires positif de détecté.

Pour visualiser la variable « Statut de survie » nous utiliserons plutôt la fonction :

```
table(data$`Statut de survie`)
barplot(table(data$`Statut de survie`),
        main = "Repartition du statut de survie", col = c("green", "red"))
```

Cette fonction nous donne ce résultat :



Nous pouvons confirmé notre constatation, réalisé grâce aux valeurs de la fonction « summary(data) », plus de 50% des patients survive au-delà de 5 ans après l'opération.

3.2. Analyse bivariée

L'analyse bivariée est une technique statistique qui permet d'étudier la relation de deux variables entre elles. Cette analyse permet de répondre aux questions :

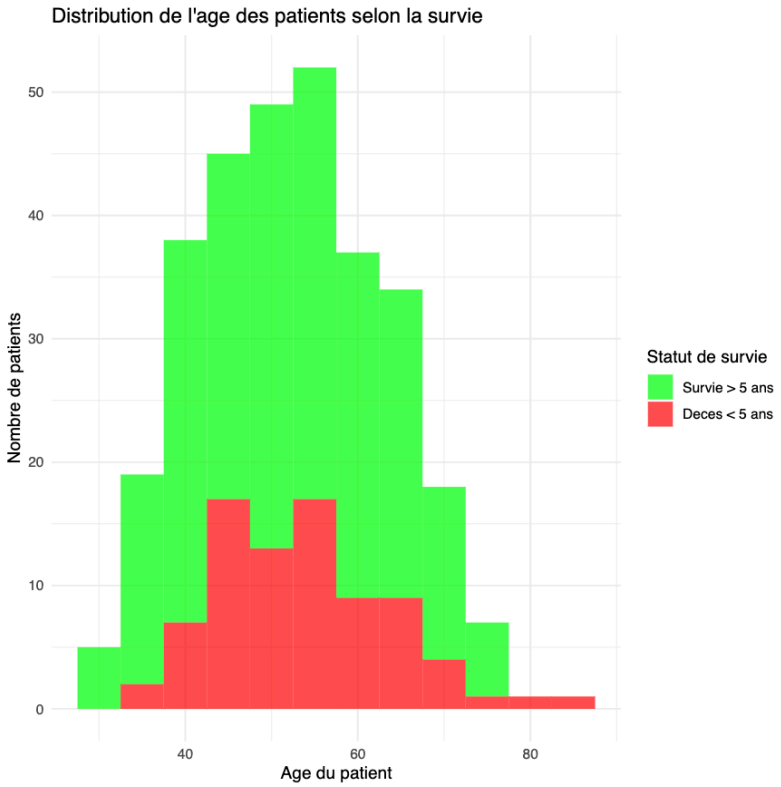
- Ces deux variables sont-elles liées ?
- L'une influence-t-elle l'autre ?
- Quelle est la force et la nature de cette relation ?

3.2.1. Age & Survie :

Nous allons, pour commencer, mettre en lien les variables représentant, l'âge des patients et la survie des patients avec la fonction :

```
ggplot(data, aes(x = `Age du patient`, fill = as.factor(`Statut de survie`))) +
  geom_histogram(binwidth = 5, position = "stack", alpha = 0.7) +
  labs(title = "Distribution de l'age des patients selon la survie",
       x = "Age du patient", y = "Nombre de patients",
       fill = "Statut de survie") +
  scale_fill_manual(values = c("green", "red"),
                    labels = c("Survie >= 5 ans", "Deces < 5 ans")) +
  theme_minimal()
```

Nous obtenons ce résultat :



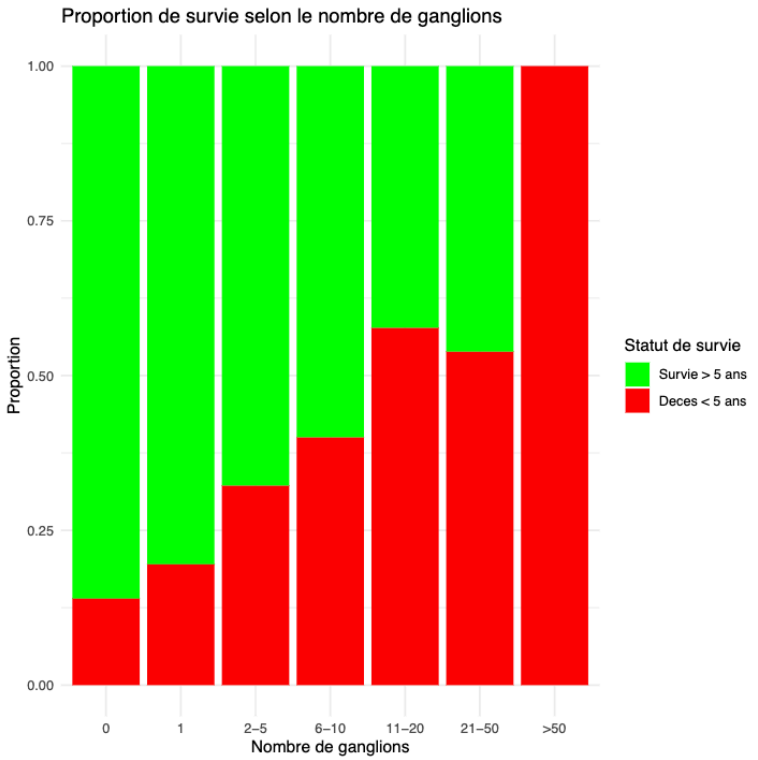
Nous pouvons observer, grâce à ce diagramme, que les patients aux alentours de 30 ans ont tous survécu plus de 5 ans après l'opération. Tandis que les patients aux alentours de 80 ans n'ont pas survécu plus de 5 ans.

3.2.2. Ganglions & Survie :

Ensuite nous pouvons mettre en lien le nombre de ganglions axillaires positifs détectés et le statut de survie. Nous avons utilisé cette fonction :

```
data$Ganglions_group <- cut(data$`Nombre de ganglions axillaires positifs`,
                            breaks = c(-1, 0, 1, 5, 10, 20, 50, Inf),
                            labels = c("0", "1", "2-5", "6-10", "11-20",
                                       "21-50", ">50"))
ggplot(data, aes(x = Ganglions_group, fill = as.factor(`Statut de survie`))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion de survie selon le nombre de ganglions",
       x = "Nombre de ganglions", y = "Proportion",
       fill = "Statut de survie") +
  scale_fill_manual(values = c("green", "red"),
                    labels = c("Survie > 5 ans", "Deces < 5 ans")) +
  theme_minimal()
```

Nous obtenons ce résultat :



Nous observons que le nombre de ganglions axillaires positifs influence grandement la survie au-delà de 5 ans après l'opération. À partir de 10 ganglions axillaires positifs détectés, il y'a 50% de change que le patient ne survivre pas plus de 5 ans après l'opération.

3.2.3. Matrice de corrélation :

Nous allons calculer la matrice de corrélation entre les trois variables quantitatives suivantes : l'âge du patient, l'année de l'opération (centrée sur 1900), et le nombre de ganglions axillaires positifs détectés et le statut de survie.

Nous avons utilisés cette fonction:

```r
# Matrice de corrélation
cor_matrix <- cor(data[, 1:4])

# On ajuste les marges pour bien voir le titre (haut = 4)
par(mar = c(1, 1, 4, 1))

# Corrplot avec paramètres d'affichage ajustés
corrplot(cor_matrix, method = "color", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         addCoef.col = "black", tl.col = "black", number.cex = 0.8,
         title = "Matrice de correlation - Haberman",
         mar = c(0, 0, 1, 0))   # nolint
```

Nous obtenons ce résultat :

L'analyse de cette matrice nous permet d'observer que :

- **Le nombre de ganglions axillaires positifs détectés** est **modérément corrélé négativement avec le statut de survie** ($r \approx 0.29$). Cela signifie que plus le nombre de ganglions axillaires positifs est élevé, plus la probabilité de ne pas survivre 5 ans après l'opération est importante.
- Cette matrice de corrélation nous permet aussi de se rendre compte que l'âge des patients n'influence la survie du patient au-delà de 5 ans après l'opération.
- Les autres variables (âge du patient, année de l'opération) présentent des **corrélations très faibles** voire nulles avec le statut de survie.
- Aucune corrélation forte n'est observée dans cette matrice ($|r| \geq 0.7$)

Ainsi, seule la variable **nombre de ganglions axillaires positifs** semble réellement informative pour expliquer la survie des patients à 5 ans. Les autres variables apportent peu d'information linéairement corrélée à la variable de survie.

3.3. Analyse multivariées :

3.3.1. ACP :

Une Analyse en Composantes Principales (ACP) est une méthode statistique et linéaire de réduction de dimension et d'exploration de données multivariées. Elle permet de résumer l'information (la variance) d'un jeu de variables originales en un nombre plus petit k de composantes principales (axes), tout en perdant le moins possible d'information. Nous allons réaliser une ACP sur trois variables quantitatives : l'âge du patient, l'année de l'opération, et le nombre de ganglions axillaires positifs détectés. L'objectif est de visualiser la répartition des patients en fonction de leur statut de survie.

Voici la fonction utilisée :

```r
data$Survie <- factor(data$`Statut de survie`, labels = c("Survie => 5 ans", "Deces < 5 ans"))


# Réalisation de l'ACP
acp <- PCA(data[, 1:3], scale.unit = TRUE, graph = FALSE)

# Graphique des individus avec ellipse de confiance
fviz_pca_ind(acp,
             geom.ind = "point",
             col.ind = data$Survie,
             palette = c("forestgreen", "firebrick"),
             addEllipses = TRUE,
             ellipse.level = 0.95,
             legend.title = "Statut de survie") +
  ggtitle("ACP des patients selon leurs caracteristiques") +
  theme_minimal()
```

Voici le résultat obtenu :

Les deux premières dimensions principales expliquent ensemble 70,2 % de la variance totale, ce qui permet une bonne représentation globale des données, bien qu'environ 30 % de l'information soit perdue. La projection des individus montre un chevauchement partiel entre les deux groupes de survie, mais aussi une certaine tendance à la séparation, ce qui indique que les variables retenues influencent partiellement la survie des patients.

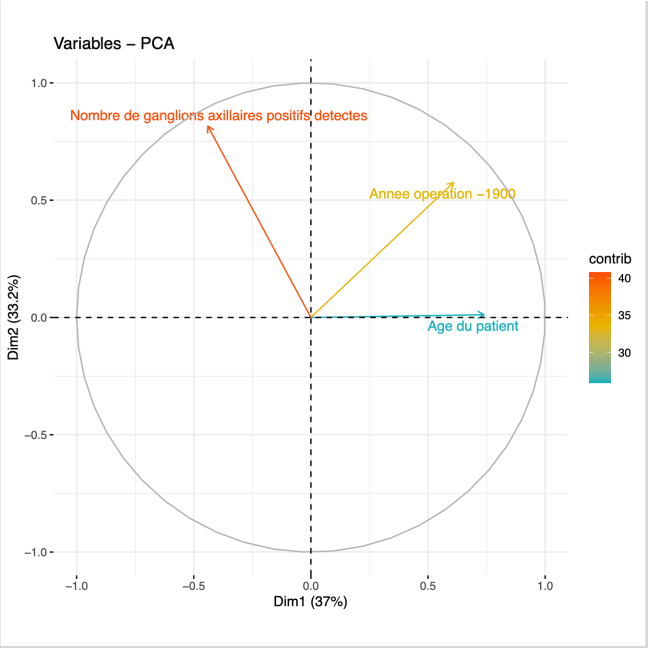
3.3.2. Cercle de corrélation :

Le cercle de corrélation (ou cercle des corrélations) est un graphique qui permet, après une ACP, de visualiser directement la relation entre chaque variable d'origine et les composantes principales (axes factoriels).

Nous avons utilisé la fonction suivante :

```
# Cercle des corrélations
fviz_pca_var(acp,
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

Voici le graphique obtenu :



On observe que :

- Le nombre de ganglions axillaires positifs est fortement corrélé avec la deuxième composante (Dim2).
- L'âge du patient est majoritairement lié à la première composante (Dim1).
- L'année de l'opération contribue modérément aux deux composantes.

L'analyse des angles entre les flèches montre que les variables sont faiblement corrélées entre elles, hormis peut-être une légère corrélation entre l'âge et l'année de l'opération.

3.3.2. K-Means clustering :

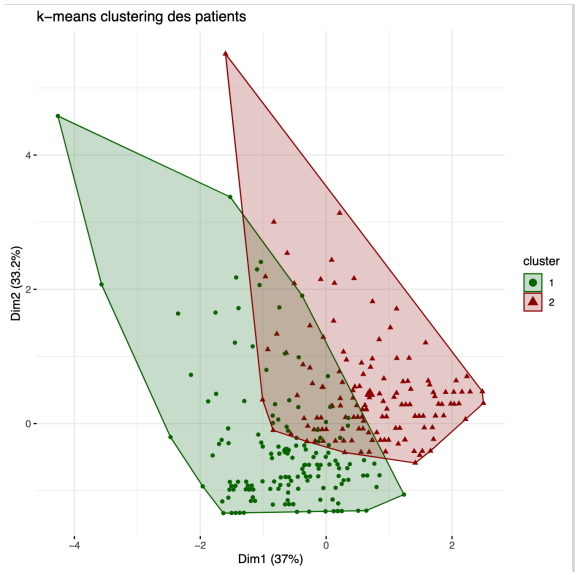
L'algorithme K-means a été appliqué pour segmenter automatiquement les patients en deux groupes (k = 2), à partir des mêmes variables quantitatives que pour l'ACP. Ce choix de k=2 repose sur le fait que le statut de survie contient deux modalités, mais aurait pu être justifié plus rigoureusement par une méthode comme la courbe du coude. Voici la fonction utilisée :

```
data_scaled <- scale(data[, 1:3])
# Calcul de la distance et clustering
d <- dist(data_scaled)
hc <- hclust(d, method = "ward.D2")


set.seed(123)
km <- kmeans(data_scaled, centers = 2, nstart = 25)

# Visualisation des clusters
fviz_cluster(km, data = data_scaled,
             geom = "point",
             palette = c("darkgreen", "darkred"),
             ggtheme = theme_minimal(),
             main = "k-means clustering des patients")
```

Voici le résultat obtenu :

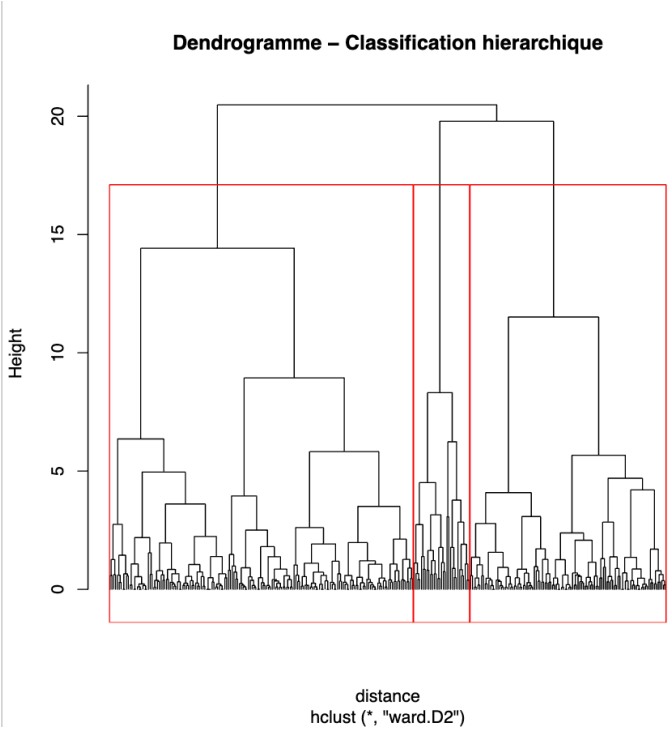


Les clusters obtenus présentent une séparation partielle, ce qui corrobore les résultats de l'ACP : les variables cliniques retenues (âge, année, ganglions) permettent de différencier partiellement les profils de patients. Il est important de noter que les étiquettes de survie n'ont pas été utilisées pour former les clusters, mais uniquement pour les colorer a posteriori.

Ainsi, bien que l'analyse multivariée ne permette pas une séparation parfaite entre les deux groupes de patients, elle met en évidence des tendances significatives et souligne l'importance du nombre de ganglions axillaires positifs dans la caractérisation du risque.

3.3.3. Classification hiérarchique :

Afin de compléter notre analyse multivariée, nous avons réalisé une classification hiérarchique ascendante (CAH) sur l'ensemble des patients, en utilisant la méthode de Ward (Ward.D2) et une matrice de distances euclidiennes.



Nous avons utilisez cette fonction :

```r
# Préparation des données (exclure la variable de survie)
data_clustering <- data[, c("Age du patient", "Annee de l'operation -1900",
                            "Nombre de ganglions axillaires positifs detectes")]

# Standardisation des données
data_clustering_scaled <- scale(data_clustering)

# Calcul de la matrice de distances
distance <- dist(data_clustering_scaled, method = "euclidean")

# Clustering hiérarchique avec la méthode de Ward
hc <- hclust(distance, method = "ward.D2")

# Affichage du dendrogramme
plot(hc, labels = FALSE, hang = -1, main = "Dendrogramme - Classification hierarchique")
rect.hclust(hc, k = 3, border = "red")  # Découper en 3 groupes, ajustable
```

Ce dendrogramme permet de visualiser la structure de regroupement naturelle des patients en fonction de leurs caractéristiques cliniques. En traçant une ligne de coupure à une hauteur appropriée, nous avons choisi de diviser l'ensemble des observations en 3 groupes principaux, comme indiqué par les rectangles rouges.

L'interprétation du dendrogramme montre que :

- Chaque groupe rassemble des patients ayant des profils relativement homogènes.
- La séparation entre les groupes semble marquée, ce qui témoigne de différences significatives entre les patients sur les variables considérées.
- Ces résultats sont cohérents avec ceux obtenus par l'ACP et le clustering K-means, et confirment l'importance de certaines variables, notamment le nombre de ganglions axillaires positifs, dans la différenciation des profils de survie.

4. Conclusion

Tout au long de ce rapport nous avons pu analyser les données d'une étude, réalisée de 1958 et 1970 à l'université de Chicago, portant sur la survie des patients ayant subi une intervention chirurgicale pour un cancer du sein. Cette étude nous à permis grave à l'étude univariée et bivariée d'exprimer de manière lisible et simple les données.

L'analyse univariée nous a permis d'identifier les distributions caractéristiques des principales variables, notamment la répartition des âges et le nombre de ganglions axillaires. Ces visualisations ont mis en évidence des tendances marquantes au sein de l'échantillon étudié, permettant ainsi de mieux comprendre la structure de nos données.

L'analyse bivariée a, quant à elle, révélé des relations intéressantes entre ces variables et le statut de survie des patients. En particulier, la comparaison des proportions de survie en fonction du nombre de ganglions axillaires a suggéré que les patients présentant un nombre réduit de ganglions axillaires atteints tendent à avoir de meilleures perspectives de survie. Ces résultats soulignent l'importance de certains facteurs cliniques dans l'évaluation du pronostic, et ouvrent la voie à des investigations complémentaires sur d'éventuelles interactions avec d'autres variables explicatives.

Enfin, l'analyse multivariée a permis d'explorer les structures globales des données à travers l'Analyse en Composantes Principales (ACP) et le clustering. L'ACP a révélé que les deux premières dimensions expliquaient plus de 70 % de la variance, offrant ainsi une bonne visualisation des individus selon leurs caractéristiques. Les ellipses de confiance ont mis en évidence une tendance à la séparation entre les groupes de survie, tandis que le cercle de corrélation a confirmé la contribution importante de certaines variables, comme le nombre de ganglions axillaires positifs.

Le clustering par K-means a également permis d'identifier deux groupes distincts de patients, en accord avec les observations de l'ACP. Ces résultats montrent que les caractéristiques cliniques des patients influencent significativement leur regroupement, et confirment la pertinence d'approches multivariées pour enrichir l'interprétation des données médicales.

La classification hiérarchique nous a apporté une validation supplémentaire sur la structure sous-jacente de nos données et renforce l'idée que les caractéristiques cliniques utilisées sont pertinentes pour discriminer les risques de survie des patients.