Programmation Logiciels Statistiques - Projet SAS

Victor Haguet, Pierre Lague - L3 CMI Université Bretagne Sud

28 février 2022

Résumé

Ce projet met en place l'analyse uni/bi/multivariée de données cardiaques récupérées sur Kaggle. Ces données sont anonymisées et comprennent des mesures telles que : trestbps (pression sanguine au repos), chol (taux de cholesterol en mg/ml), thalach (rythme cardiaque maximal), oldpeak (rythme cardique après effort), target (0, 1 indique la présence d'une maladie cardiaque ou non), age (âge de l'individu). Les langages utilisés dans l'analyse de ses données sont : SAS, R et Python (notebook python fourni avec l'archive de projet). La plupart des graphes seront ceux obtenus sous R (librairie ggplot) et sous SAS. Leur interprétation sera issue des résultats obtenus sous SAS et R. Nous tenterons de répondre à la problématique suivante : "Quelles caractéristique physiologique et cardiaque pouvons nous utiliser pour expliquer la présence de maladie cardiaque chez un individu ?".

Table des matières

1 Introduction

Il est évident que notre objectif est d'expliquer la variable qualitative discrète "target" à l'aide des autres variables explicatives (chol, thalach, oldpeak, trestbps, age). Le but est de savoir si les variables explicatives jouent un rôle dans la présence, ou non, d'une maladie cardiaque chez l'individu observé. Nous mènerons donc des analyses statistiques pour tenter de trouver des relations entre les variables et expliquer la variable "target".

	age	trestbps	chol	thalach	oldpeak	target
1	63	145	233	150	2.3	1
2	37	130	250	187	3.5	1
3	41	130	204	172	1.4	1
4	56	120	236	178	0.8	1
5	57	120	354	163	0.6	1
6	57	140	192	148	0.4	1
7	56	140	294	153	1.3	1
8	44	120	263	173	0.0	1
9	52	172	199	162	0.5	1
10	57	150	168	174	1.6	1

FIGURE 1 – Echantillon des données d'étude

2 Analyse Univariée

Dans cette section nous allons nous concentrer sur l'analyse univariée des données. Nous allons donc tenter de déterminer comment les données sont distribuées, comment sont réparties leurs valeurs en émmettant des hypothèses et en les validant, ou non à l'aide de tests statistique de normalité. Entre autres, ceux de Kolomogorov-Smirnov, Shapiro-Wilk, Cramer-von Mises et Anderson-Darling.

2.1 Gestion des données et outliers

La première étape d'une analyse de données est une étape appelée "data engineering" qui à pour but de vérifier l'intégrité des données. Dans cette étape nous vérifions le nombre d'outliers et leurs valeurs a l'aide de boxplot puis nous supprimerons les tuples de données possédant des valeurs trop élevées.

Ci-dessous, le résultat d'un programme sur R(ggplot2) qui définit et trace les boxplots pour chacune des variables explicatives. L'équivalent à été réalisé sous SAS avec un macro programme.

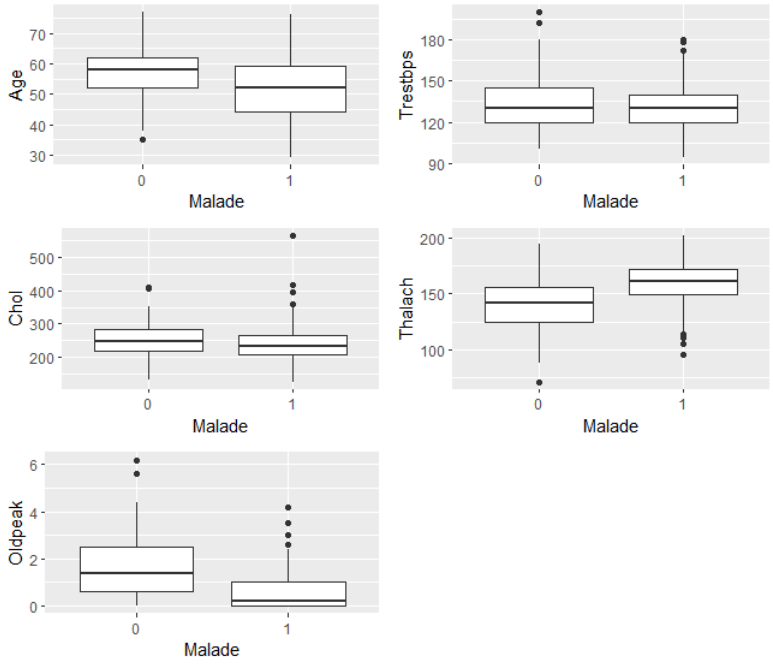


FIGURE 2 – Boxplots : répartition des valeurs des variables explicatives en fonction de target

(a) Il y a des outliers présents pour toutes les variables.

En prenant en compte les résultats donnés nous avons donc effectué des modifications sur les données en supprimant des données supérieures ou inférieures à un certain seuil dans les colonnes thalach, oldpeak, chol et trestbps.

Maintenant que nos données ont été arrangées de sorte que les outliers n'influencent pas les tests statistiques de l'analyse univariée, nous pouvons nous pencher sur notre première variable explicative discrète : "age".

2.2 Analyse de l'âge

Ici, l'analyse sera basée en priorité sur le test de Shapiro-Wild car celui-ci est moins exigeant que les autres tests. Ce test a put être réalisé sur SAS et sur R contrairement aux autres tests qui ne retournent pas les mêmes P-value suivant le language utilisé. Il semblerait que les résultats obtenus par SAS soient les plus cohérents. Sous SAS nous pouvons utiliser la proc univariate, means et npar1way afin de déterminer la distribution de la variable age en émettant les hyptohèses de normalité.

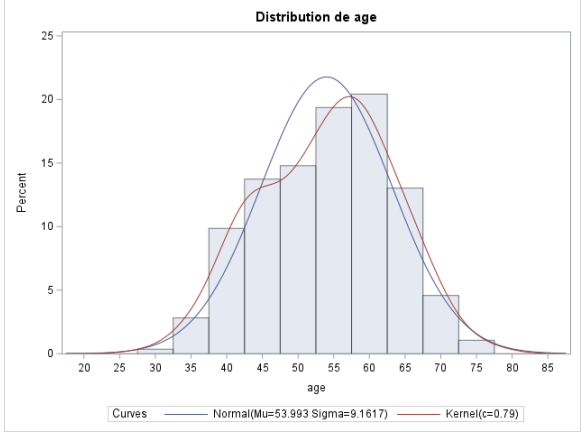
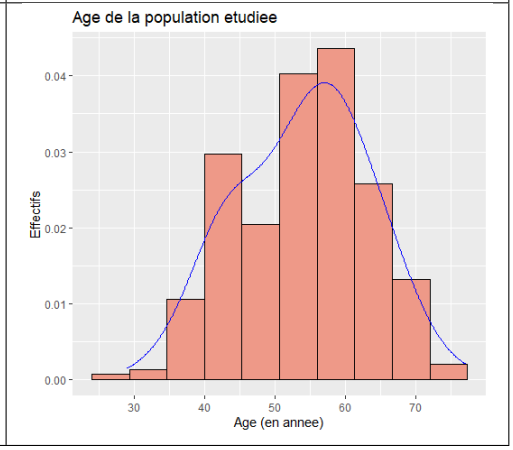
Distribution de l'age sur SAS	Distribution de l'age sur R
	

TABLE 1 – Distribution de la variable age sur SAS et R

Nous pouvons observer sur les graphes précédents que la distribution de la variable "age" est bimodale. Ceci indique qu'elle ne suit pas une loi normale.

Plusieurs tests statistiques de normalité peuvent alors être réalisés.

Tests de normalité				
Test	Statistique		p-value	
Shapiro-Wilk	W	0.987602	Pr < W	0.0154
Kolmogorov-Smirnov	D	0.079331	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.231568	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.323846	Pr > A-Sq	<0.0050

FIGURE 4 – Les tests satistiques de normalité rejètent l'hypothèse de normalité

Ici, les Test de Shapiro-Wilk, Kolmogorov-Smirnov, de Cramer-von Mises et de Anderson-Darling ont été réalisés et chacun d'eux ont rejeté l'hypothèse de normalité. La p-value (que nous appelerons ϕ) est inférieur au risque d'erreur $\alpha = 5\%$. On en conclue donc que la variable age ne suit pas une loi normale.

2.3 Analyse univariée des autres variables explicatives

En suivant une procédure équivalente pour l'analyse univariée de la variable "age" nous pouvons dresser les tableaux suivants qui résument l'analyse univariée de chaque variable avec les résultats des tests statistiques (acceptation ou rejet de l'hypothèse de normalité). Puis leur distribution (comparaison au noyau de la loi normale) afin d'illustrer les résultats des test précédents.

Ci-dessous le tableau résumant les valeurs de chaque test statistique de normalité et, respectivement, la validation ou non de l'hyptohèse de normalité :

Variables / Test	thalach	trestbps	chol	oldpeak
Shapiro-Wilk	$\phi < 0.0001$	$\phi < 0.033$	$\phi > 0.1910$	$\phi < 0.0001$
Kolmogorov-Smirnov	$\phi < 0.010$	$\phi < 0.010$	$\phi > 0.150$	$\phi < 0.010$
Cramer-Von Mises	$\phi < 0.005$	$\phi < 0.005$	$\phi \sim 0.107$	$\phi < 0.005$
Anderson-Darling	$\phi < 0.005$	$\phi < 0.005$	$\phi \sim 0.098$	$\phi < 0.005$
Hyptohèse Normalité	Rejetée	Rejetée	Acceptée	Rejetée

TABLE 2 – Tableau des tests statistiques

Ci-dessous les graphes représentant la distribution des valeurs des variables explicatives thalach et trestbps :

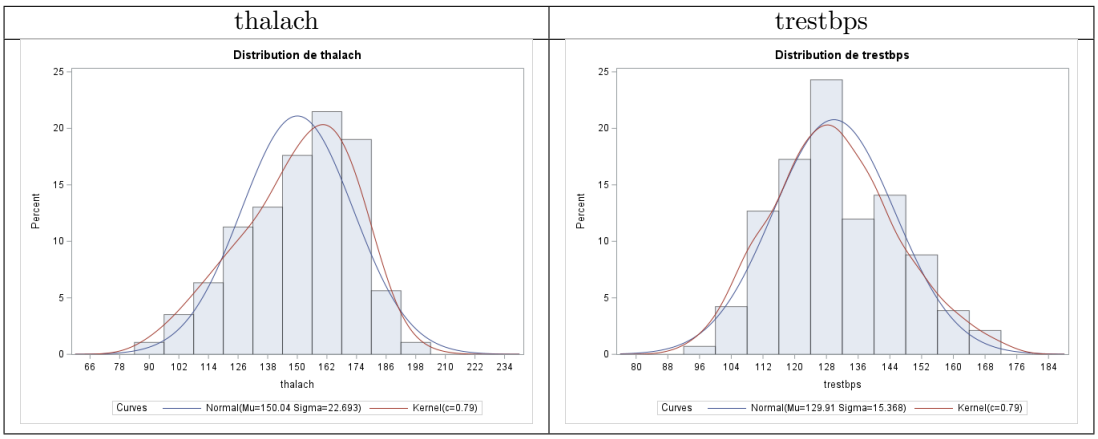


TABLE 3 – Les variables thalach et trestbps n'ont pas une distribution normale sous SAS

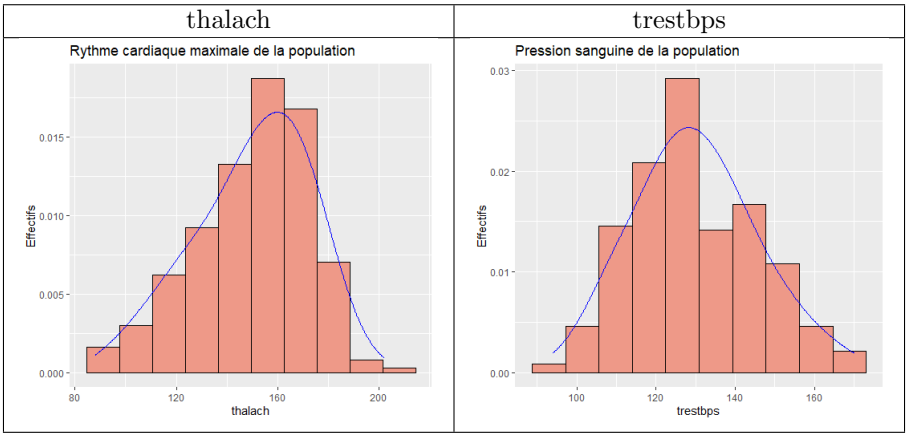


TABLE 4 – Les variables thalach et trestbps n'ont pas une distribution normale sous R

Selon le tableau 2 nous pouvons voir que les hypothèses de normalité sont rejetées pour les variables thalach et trestbps. Visuellement, il semble que leur distribution ne soit pas gaussienne (tables 3 et 4).

Ci-dessous la liste des graphes représentant la distribution des valeurs des variables explicatives chol et oldpeak :

Selon le tableau 2 nous pouvons voir que les hypothèses de normalité sont rejetées pour la variable oldpeak. Sa distribution n'est pas gaussienne selon les tableaux 5 et 6. Cependant, selon le tableau 2, on constate que la variable chol suit effectivement une loi normale et que ses valeurs ont une distribution gaussienne ce qui est logiquement observable sur les graphes de la table 5 et 6.

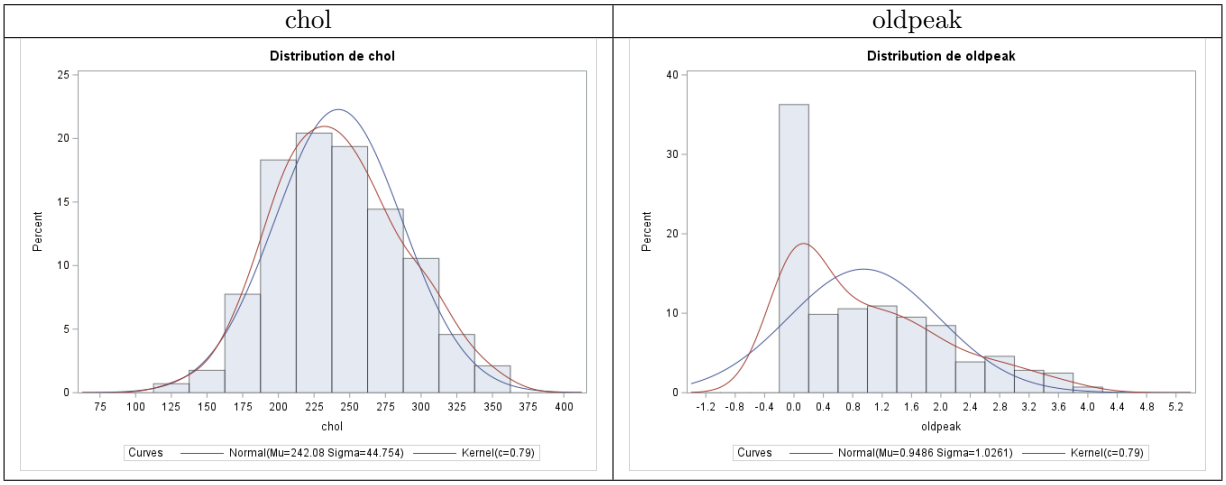


TABLE 5 – La variable chol à une distribution normale, la variable oldpeak n'a pas une distribution normale sous SAS.

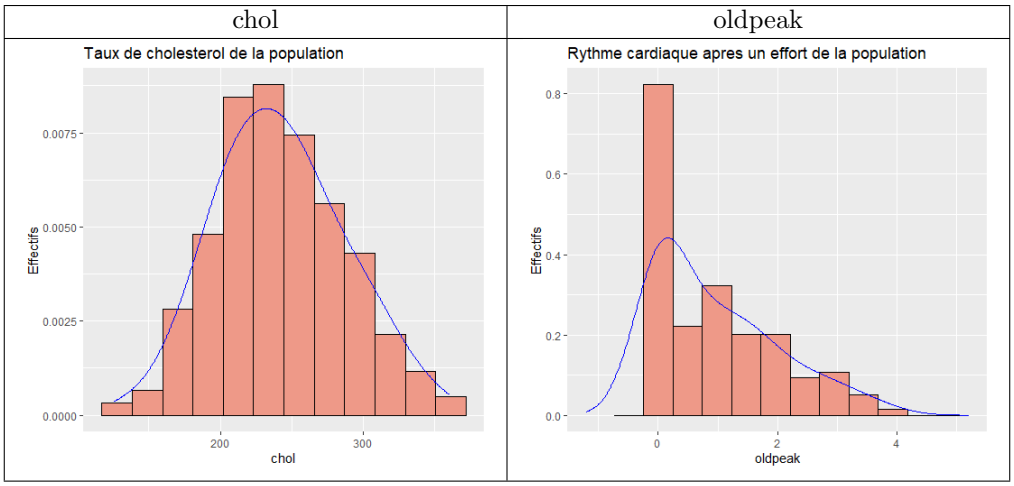


TABLE 6 – La variable chol à une distribution normale, la variable oldpeak n'a pas une distribution normale (R).

2.4 Conclusion de l'analyse univariée

Au terme de cette première analyse univariée sur les variables explicatives, nous avons pu mener une analyse descriptive des données en réduisant le nombre d'outliers pour éviter un biais trop important lors de la réalisation des tests statistiques. Nous avons également trouvé qu'une seule (chol) des cinq variables explicatives suit une loi normale. Les autres ont une distribution bimodale. Ces résultats ont été obtenus graçe à la proc univariate sous SAS en comparant la distribution des valeurs des variables sous forme d'histogramme, au noyau de la loi normale.

3 Analyse Bivariée

Dans cette partie nous mènerons une analyse bivariée sur plusieurs couples de variables. L'analyse bivariée a pour but de mettre en relation deux informations différentes afin de vérifier si elles sont liées ou non. La mise en évidence d'un tel lien peut notamment servir à faire des prédictions sur la valeur prise par une variable en fonction de celle prise par une autre. L'analyse bivariée donne aussi des informations utiles permettant ensuite de construire un modèle multivarié. Un moyen de trouver des relations entre les variables et de voir comment elles s'influencent. En utilisant la proc logistic et corr sous SAS nous allons pouvoir observer comment une variable explicative va influencer notre variable cible "target".

3.1 Analyse de la corrélation entre les variables explicatives

Cette analyse nous permet d'avoir une première idée de comment les variables vont pouvoir s'influencer les unes entre les autres. La corrélation est très souvent réduite à la corrélation linéaire entre variables quantitatives, c'est-à-dire l'ajustement d'une variable par rapport à l'autre par une relation affine obtenue par régression linéaire. Pour cela, on calcule un coefficient de corrélation linéaire, quotient de leur covariance par le produit de leurs écarts types :

Coefficients de corrélation de Pearson, N = 284
Proba > |r| sous H0: Rho=0

	age	trestbps	chol	thalach	oldpeak	target
age	1.00000	0.27774 <.0001	0.17869 0.0025	-0.41440 <.0001	0.21353 0.0003	-0.22638 0.0001
trestbps	0.27774 <.0001	1.00000	0.11183 0.0598	-0.06955 0.2427	0.14242 0.0163	-0.11375 0.0555
chol	0.17869 0.0025	0.11183 0.0598	1.00000	-0.02539 0.6701	-0.00457 0.9389	-0.10976 0.0647
thalach	-0.41440 <.0001	-0.06955 0.2427	-0.02539 0.6701	1.00000	-0.34337 <.0001	0.42438 <.0001
oldpeak	0.21353 0.0003	0.14242 0.0163	-0.00457 0.9389	-0.34337 <.0001	1.00000	-0.43575 <.0001
target	-0.22638 0.0001	-0.11375 0.0555	-0.10976 0.0647	0.42438 <.0001	-0.43575 <.0001	1.00000

FIGURE 5 – La proc corr sous SAS nous permet d'avoir les coefficients de corrélation entre chaque variables

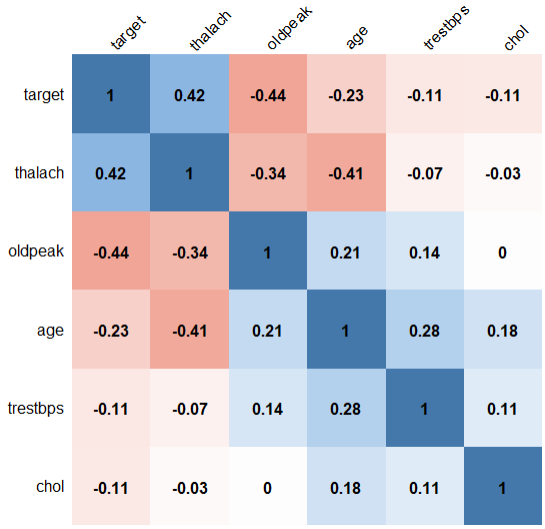


FIGURE 6 – Matrice de corrélation obtenue sous R

D'après les résultats de SAS et R les coefficients de corrélation avec target sont tous relativement faibles. Cependant on constate que :

— thalach et target sont moyennement corrélées positivement

— oldpeak et target sont moyennement corrélées négativement

3.2 Analyse de la régression logistique des variables explicatives sur target

L'analyse de la regression logistique à été choisie car notre variable cible (dependante) est catégorique (1 ou 0, repectivement, presence de maladie ou non). Le but de cette analyse est de déterminer quelle variable à un apport statistique au model de regression logistique en déterminant le paramètre $\hat{\beta}$ de l'équation du modèle ajusté. Le signe de ce paramètre indiquera son influence :

1. si $\hat{\beta} > 0$, alors si la variable explicative augmente, la probabilité que l'individu soit malade va augmenter
2. si $\hat{\beta} < 0$, alors si la variable explicative diminue, la probabilité que l'individu soit malade va diminuer

Les figures suivantes illustrent les résultats de la proc logistic sur target en fonction de chaque variable explicative. NB. Dans la proc logistic, la référence pour target est : target = 1.

3.2.1 Regression logistique de la variable target sur la variable age

Sous SAS, la sortie de la procédure est la suivante :

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	14.9100	1	0.0001
Score	14.5547	1	0.0001
Wald	13.9453	1	0.0002

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	3.0779	0.7735	15.8325	<.0001	21.714
age	1	-0.0523	0.0140	13.9453	0.0002	0.949

FIGURE 7 – Sortie de la proc logistic de target en fonction de age sous SAS

Nous constatons les particularités suivantes :

1. Les estimations des paramètres du modèle sont les suivantes $\hat{\beta}_0$ = 3.078 (target) et $\hat{\beta}_j$ = -0.052 (age). Le modèle s'écrit donc de la façon suivante : P(target=1 | age = $\hat{\beta}_j$) = 3.078 - 0.052_{age}.
2. La valeur-p pour $\hat{\beta}_j$ (Pr > Khi-2), correspondant aux test des hypothèses $H_0 : \hat{\beta}_j = 0$ versus $H_1 : \hat{\beta}_j ! = 0$, est plus grande que $10\exp(-4)$ (0.0002) et donc l'effet de la variable âge est statistiquement proche de zéro. Son influence sera donc négligeable sur l'évolution de la probabilité qu'une personne soit malade.

Sous R, on obtient les résultats suivants :

```
Call:
glm(formula = target ~ age, family = "binomial", data = heart,
    start = c(0, 0), control = list(maxit = 1000, trace = TRUE,
        epsilon = 1e-16))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7323  -1.1745   0.8141   1.0501   1.5735

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.07795    0.77355   3.979 6.92e-05 ***
age         -0.05228    0.01400  -3.734 0.000188 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 8 – Description du modèle GLM de target en fonction de age sous R

En utilisant la fonction glm (generalized linear models) sous R, on peut réaliser la régression logistique de target en fonction de age. On constate qu'on obtient les mêmes résultats que sous SAS, les coefficients sont légèrement différents mais l'impact de la variable age sur la variable target ne change pas.

3.2.2 Regression logistique de la variable target sur la variable thalach

Sous SAS, la sortie de la procédure est la suivante :

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	54.8502	1	<.0001
Score	51.1481	1	<.0001
Wald	43.1345	1	<.0001

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	-6.3381	1.0111	39.2937	<.0001	0.002
thalach	1	0.0440	0.00670	43.1345	<.0001	1.045

FIGURE 9 – Sortie de la proc logistic de target en fonction de thalach sous SAS

Nous constatons les particularités suivantes :

1. Les estimations des paramètres du modèle sont les suivantes $\hat{\beta}_0$ = -6.338 (target) et $\hat{\beta}_j$ = 0.044 (thalach). Le modèle s'écrit donc de la façon suivante : P(target=1 | thalach = $\hat{\beta}_j$) = -6.338 + 0.044_{thalach}.
2. La valeur-p pour $\hat{\beta}_j$ (Pr > Khi-2), correspondant aux test des hypothèses $H_0 : \hat{\beta}_j = 0$ versus $H_1 : \hat{\beta}_j ! = 0$, est plus petite que $10\exp(-4)$ et donc l'effet de la variable thalach est statistiquement différent de zéro. Plus thalach va augmenter plus la probabilité qu'un individu soit malade va augmenter.

En utilisant la fonction glm (generalized linear models) sous R, on peut réaliser la régression logistique de target en fonction de thalach. On constate qu'on obtient les mêmes résultats que sous SAS. On retrouve bien les coefficients suivants : $\hat{\beta}_0$ = -6.338 (target) et $\hat{\beta}_j$ = 0.044 (thalach). On peut en conclure que l'effet de la variable thalach reste le même sur la variable target :

```
Call:
glm(formula = target ~ thalach, family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9192  -1.0398   0.6179   0.9017   2.1112

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.338220   1.011113  -6.269 3.64e-10 ***
thalach      0.043996   0.006699   6.568 5.11e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 10 – Description du modèle GLM de target en fonction de thalach sous R

3.2.3 Regression logistique de la variable target sur la variable chol

La sortie de la procédure est la suivante :

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	3.4343	1	0.0639
Score	3.4217	1	0.0643
Wald	3.3845	1	0.0658

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	1.4515	0.6704	4.6869	0.0304	4.269
chol	1	-0.00499	0.00271	3.3845	0.0658	0.995

FIGURE 11 – Sortie de la proc logistic de target en fonction de chol sous SAS

Nous constatons les particularités suivantes :

1. Les estimations des paramètres du modèle sont les suivantes $\hat{\beta}_0 = 1.452$ (target) et $\hat{\beta}_j$ = -0.005 (chol). Le modèle s'écrit donc de la façon suivante : P(target=1 | chol = $\hat{\beta}_j$) = 1.452 - 0.005_{chol}.
2. La valeur-p pour $\hat{\beta}_j$ (Pr > Khi-2), correspondant aux test des hypothèses $H_0 : \hat{\beta}_j = 0$ versus $H_1 : \hat{\beta}_j ! = 0$, est plus grande que $10\exp(-4)$ et donc l'effet de la variable chol est statistiquement nul. Donc la variable chol n'influencera pas la probabilité qu'une personne soit malade.

En utilisant la fonction glm (generalized linear models) sous R, on peut réaliser la régression logistique de target en fonction de chol. De même que pour la variable age, l'effet de chol sur la variable target est statistiquement nul :

```
Call:
glm(formula = target ~ chol, family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5294  -1.2526   0.9763   1.0702   1.3271

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.451460   0.670439   2.165   0.0304 *
chol        -0.004991   0.002713  -1.840   0.0658 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 12 – Description du modèle GLM de target en fonction de chol sous R

3.2.4 Regression logistique de la variable target sur la variable oldpeak

La sortie de la procédure est la suivante :

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	57.6234	1	<.0001
Score	53.9249	1	<.0001
Wald	44.7135	1	<.0001

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	1.1831	0.1886	39.3601	<.0001	3.265
oldpeak	1	-0.9995	0.1495	44.7135	<.0001	0.368

FIGURE 13 – Sortie de la proc logistic de target en fonction de oldpeak sous SAS

Nous constatons les particularités suivantes :

1. Les estimations des paramètres du modèle sont les suivantes $\hat{\beta}_0 = 1.183$ (target) et $\hat{\beta}_j$ = -0.999 (oldpeak). Le modèle s'écrit donc de la façon suivante : P(target=1 | oldpeak = $\hat{\beta}_j$) = 1.183 - 0.999_{oldpeak}.
2. La valeur-p pour $\hat{\beta}_j$ (Pr > Khi-2), correspondant aux test des hypothèses $H_0 : \hat{\beta}_j = 0$ versus $H_1 : \hat{\beta}_j ! = 0$, est plus petite que $10\exp(-4)$ et donc l'effet de la variable oldpeak est statistiquement différent de zéro. Plus oldpeak va diminuer, plus la probabilité qu'une personne soit malade va diminuer.

Sous R, on retrouve les mêmes coefficients du modèle de régression logistique de target en fonction de oldpeak :

```
Call:
glm(formula = target ~ oldpeak, family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7031  -1.0066   0.7310   0.8408   2.1951

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1831     0.1886   6.274 3.52e-10 ***
oldpeak      -0.9995     0.1495  -6.687 2.28e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 14 – Description du modèle GLM de target en fonction de oldpeak sous R

3.2.5 Regression logistique de la variable target sur la variable trestbps

La sortie de la procédure est la suivante :

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	3.6882	1	0.0548
Score	3.6744	1	0.0553
Wald	3.6300	1	0.0567

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	2.2008	1.0372	4.5021	0.0339	9.032
trestbps	1	-0.0151	0.00791	3.6300	0.0567	0.985

FIGURE 15 – Sortie de la proc logistic de target en fonction de trestbps sous SAS

Nous constatons les particularités suivantes :

1. Les estimations des paramètres du modèle sont les suivantes $\hat{\beta}_0 = 2.201$ (target) et $\hat{\beta}_j$ = -0.015 (trestbps). Le modèle s'écrit donc de la façon suivante : P(target=1 | trestbps = $\hat{\beta}_j$) = 2.201 - 0.015_{trestbps}.
2. La valeur-p pour $\hat{\beta}_j$ (Pr > Khi-2), correspondant aux test des hypothèses $H_0 : \hat{\beta}_j = 0$ versus $H_1 : \hat{\beta}_j ! = 0$, est plus grande que $10\exp(-4)$ et donc l'effet de la variable oldpeak est statistiquement nul. La variable trestbps ne va donc pas influencer la probabilité qu'une personne soit malade.

Sous R, on retrouve les mêmes coefficients du modèle de régression logistique de target en fonction de trestbps. Tout comme age et chol, la variable trestbps ne permet pas d'expliquer la variable target.

```
Call:
glm(formula = target ~ trestbps, family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4828  -1.2687   0.9569   1.0764   1.3338

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.200829   1.037230   2.122   0.0339 *
trestbps    -0.015066   0.007908  -1.905   0.0567 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 16 – Description du modèle GLM de target en fonction de trestbps sous R

3.3 Conclusion de l'analyse bivariée

Au terme de cette analyse nous avons constaté les points suivants :

1. Sous R ou sous SAS nous trouvons les mêmes résultats quand aux capacités explicatives de chaque variable.
2. Les variables étant moyennement correlées avec target sont : oldpeak et thalach. Les autres variables ne sont que très peux correllées avec target.
3. A l'aide de la proc logistic nous pouvons affirmer que :
   (a) thalach va influencer positivement la probabilité qu'une personne soit malade
   (b) oldpeak va influencer négativement la probabilité qu'une personne soit malade
   (c) les variables chol, age et trestbps n'influencent pas (à elles seules) la probabilité qu'une personne soit malade.

Dans les sorties SAS le tableau "Test de l'hypothèse nulle globale : $\beta = 0$" contient les résultats de trois tests pour l'hypothèse nulle que tous les paramètres sont nuls, contre l'alternative qu'au moins un des paramètres est différent de zéro. Comme il y a un seul paramètre ici, ces tests reviennent à tester l'effet de la variable en question.
On retient donc que les variables thalach et oldpeak sont les seules qui jouent effectivement un rôle dans l'explication de la variable target.

4 Analyse Multivariée

L'analyse multivariée regroupe les méthodes statistiques qui s'attachent à l'observation et au traitement simultané de plusieurs variables statistiques en vue d'en dégager une information synthétique pertinente. Les deux grandes catégories de méthodes d'analyse statistique multivariées sont, d'une part, les méthodes dites descriptives et, d'autre part, les méthodes dites explicatives. Les méthodes descriptives ont pour objectif d'aider à structurer et résumer un ensemble de données issues de plusieurs variables, sans privilégier particulièrement l'une de ces variables. Toutes les variables sont donc prises en compte au même niveau. Les traitements et représentations graphiques visent à apporter une vision globale la plus exacte possible de l'ensemble des données analysées, en minimisant la déperdition d'information.

Les méthodes descriptives d'analyse multivariée qui seront utilisées dans cette section sont :

1. l'ACP ou analyse en composantes principales
2. le clustering

Les méthodes explicatives ont, quant à elles, pour objectif d'expliquer l'une des variables (dite dépendante) à l'aide de deux ou plusieurs variables explicatives (dites indépendantes). La méthode explicative utilisée dans cette section est :

1. la régression logistique

4.1 Analyse en composantes principales (ACP)

L'ACP a pour but de décorréler les variables quantitatives de l'étude. L'idée sera de réduire le jeu de variable à un nombre minime de variables appelées "composantes principales" afin de simplifier le jeu de donnée tout en conservant le plus d'observations.

Ici, il sera choisit de sélectionner 3 composantes principales qui seront crées à partir des variables précédemment utilisées telles que age, trestbps, chol, thalach et oldpeak. Les composantes seront ainsi exprimées en fonction de ces variables.

4.1.1 Sous SAS

A l'aide de la proc princomp sous SAS, il est donc possible de représenter les caractéristiques de ces composantes principales sous forme de graphe.

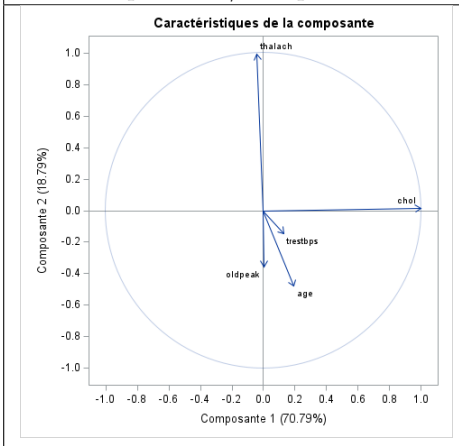
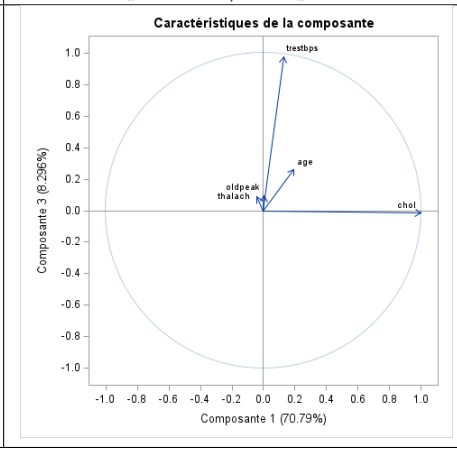
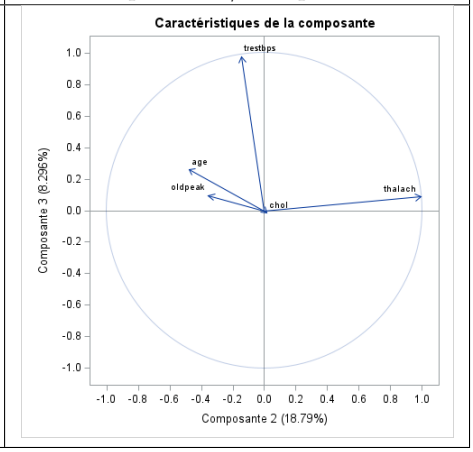
Composante1/Composante2	Composante1/Composante3	Composante2/Composante3
		

TABLE 7 – Caractéristiques des trois composantes principales

Il est possible de réaliser plusieurs déductions à partir de ces graphes :

1. Tout d'abord, on peut voir que la première composante explique 70.79% de la variation totale tandis que la seconde composante correspond à 18.79% de la variation totale et que la troisième ne correspond qu'à 8.296% de la variation totale.
2. Ensuite, il est possible de regarder plus en détail les caractéristiques de chaques composantes. La première composante oppose age, chol, trestbps et oldpeak à thalach et la seconde oppose chol et thalach à age, trestbps et oldpeak. La variable qui affecte le plus la première composante est chol tandis que la variable qui affecte le plus la seconde est thalach. La variable trestbps est celle qui impacte le plus la trosième composante mais comme le pourcentage de variation totale de celle-ci est très faible, son influence est negligeable.

De ces observations, on peut noter une grande importance de la variable chol dans les variations totale car c'est la variable la plus impactante de la première composante qui est de loin la composante la plus influente dans les variation. La variable thalach semble également avoir un réel intérêt étant la plus impactante dans la seconde composante.

4.1.2 Sous R

Pour la méthode sur R, cinq composantes principales apparaissent. Comme dit précédemment, seuls les trois composantes les plus impactantes sur la variable target seront étudiées.

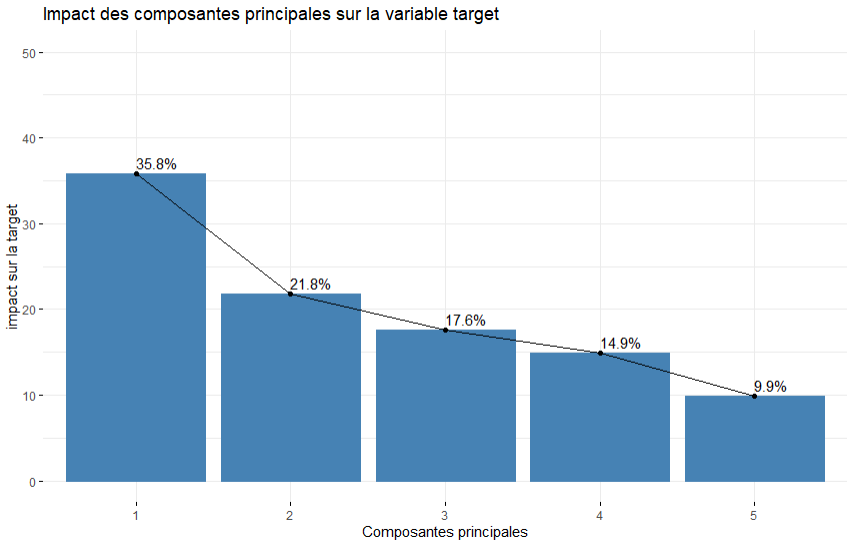


FIGURE 17 – Impact des composantes principales sur la variable target

Il est alors intéressant d'observer les variables qui impactent le plus ces trois composantes principales. Ces observations seront réalisées à l'aide des graphes ci-dessous :

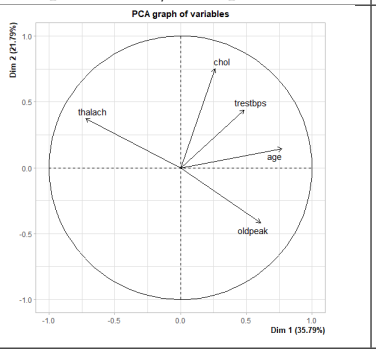
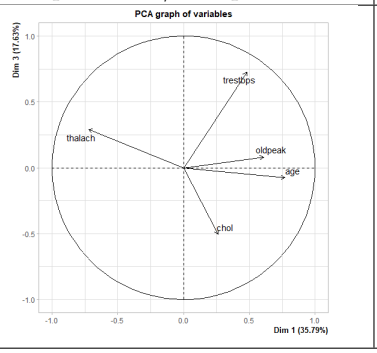
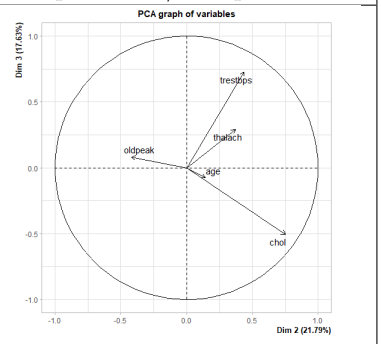
Composante 1/Composante 2	Composante 1/Composante 3	Composante 2/Composante 3
		

TABLE 8 – Composantes principales sous R

Il est possible de réaliser plusieurs déductions à partir de ces graphes :

1. Tout d'abord, nous pouvons observer que la première composante explique 35.8% de la variation totale tandis que la seconde composante correspond à 21.8% de la variation totale et que la troisième correspond à 17.8% de la variation totale.
2. Ensuite, il est possible de regarder plus en détail les caractéristiques de chaques composantes. La première composante oppose age, chol, trestbps et oldpeak à thalach et la seconde oppose chol, thalach, age et trestbps à oldpeak. La troisième composante oppose chol et age à trestbps, thalach et oldepak. Les variables qui affectent le plus la première composante sont age et thalach et que la variable qui affecte le plus la seconde est chol. La variable trestbps est celle qui impacte le plus la trosième composante mais comme le pourcentage de variation totale de celle-ci est faible, son influence est negligeable.

Ces résultats se traduisent par le graphe ci-dessous :

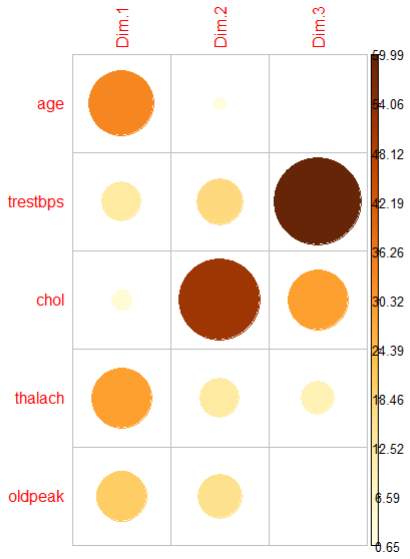


FIGURE 18 – Impact des variables explicatives sur les différentes composantes

Nous décidons alors d'exclure les variables trestbps et age de l'analyse, étant donné qu'elles n'influencent pas les composantes principales. De ce fait, à l'échelle des données entières, elle seront négligeables.

4.2 Régression logistique des combinaison de variables

Comme expliqué à la section 8 nous allons effectuer une regression logistique sur la variable target en fonction d'une combinaison de nos variables explicatives. Nous ne garderons que le variables chol, thalach et oldpeak. Nous avons vu dans la section précédente (Analyse Composante Principales) que les variables age et trestbps n'influencait pas l'explication de la variable target. Des conclusions similaires ont étés tirées de la section 12.

Nous allons d'abord tenter de trouver l'équation du modèle de regression logistique de target en fonction de la combinaison des 3 variables chol, thalach et oldpeak. Puis nous déterminerons le rapport de cote et leur intervalles de confiance respectifs pour déterminer l'effet qu'a chaque variable dans l'équation :

1. Si le rapport de cotes est supérieur à 1, alors la variable étudiée sera proportionnelle au risque que l'individu soit malade, c'est à dire que target = 1.
2. Si le rapport de cotes est inférieur à 1, alors la variable étudiée sera inversement propotionnelle au risque que l'individu soit malade, target = 0

Ci-dessous le résultat de la proc logisitc sous SAS :

Valeurs estimées du paramètre et intervalle de confiance de vraisemblance de profil			
Paramètre	Estimation	Intervalle de confiance à95%	
Intercept	2.5892	0.0964	5.1567
chol	0.00717	0.000838	0.0137
thalach	-0.0358	-0.0502	-0.0223
oldpeak	0.8079	0.5183	1.1193

FIGURE 19 – Extrait de la sortie de la proc logistic sous SAS

Ci-dessous une description du modèle de regression logistique sous R :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.589220   1.285208   2.015   0.0439 *
chol         0.007169   0.003269   2.193   0.0283 *
thalach     -0.035806   0.007100  -5.043 4.58e-07 ***
oldpeak      0.807949   0.152692   5.291 1.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 20 – Extrait de la description du modèle de régression logistique sous R

Les résultats sont les mêmes sous R et sous SAS. Nous pouvons observer que :

1. Les coefficients du modèle de régression sont les suivants :
   $\hat{\beta}_0 = 2.589$ (target), $\hat{\beta}_1 = 0.007$ (chol), $\hat{\beta}_2 = -0.036$ (thalach), $\hat{\beta}_3 = 0.808$ (oldpeak).
2. On obtient l'équation de modèle suivant : $P(\text{target}=1 \mid X=x) = 2.589 + 0.007_{\text{chol}} - 0.036_{\text{thalach}} + 0.808_{\text{oldpeak}}$
3. La variable chol influence positivement la probabilité qu'une personne soit malade (de façon presque négligeable).
4. La variable thalach influence négativement la probabilité qu'une personne soit malade.
5. La variable oldpeak influence positivement la probabilité qu'une personne soit malade (sont influence est la plus importante).

A présent passons à l'analyse du rapport de cote de chaque variable dans le modèle de regression logistique. Ci-dessous ce trouvent les rapports de cotes obtenu grâce à la régression logistique multivariée :

Estimation du rapport de cotes			
Effet	Estimation du point	Intervalle de confiance de Wald à95%	
chol	0.993	0.987	0.999
thalach	1.036	1.022	1.051
oldpeak	0.446	0.330	0.601

FIGURE 21 – Rapports de cotes des variables chol, thalach et oldpeak et IC respectifs sous SAS

De cette table, on peut conclure que plus le taux de cholestérol est faible, plus il y a de chance que l'individu soit malade. Il en va de même pour le rythme cardiaque après un effort (oldpeak). Par contre, plus le rythme cardiaque maximal est élevé, plus il y a de chance que l'individu soit malade. Il est également notable que les variations enclenchées par la variable chol sont bien moins impactantes que celles enclenchées par les variables thalach et oldpeak. Sous R nous obtenons les résultats suivants :

```
                OR       2.5 %    97.5 %
(Intercept) 0.07507855 0.005759882 0.9081346
chol        0.99285670 0.986394994 0.9991648
thalach     1.03645477 1.022547383 1.0515107
oldpeak     0.44577154 0.326508255 0.5955431
```

FIGURE 22 – Rapports de cotes des variables chol, thalach et oldpeak et IC respectifs sous R

Pour les variables chol et thalach, le point du rapport de cote est le même à 10e-3 près, de même pour les bornes de l'intervalle de confiance. Cependant pour la variable oldpeak, la borne inférieur de l'intervalle de confiance est légèrement inférieur à ce qui est obtenu sous SAS. Ceci peut être lié à la différence d'algorithme d'optimisation utilisés.

4.2.1 Conclusion de la regression logisitque multivariée

Au terme de cette étude multivariée sur la regression logistique de target en fonction de triplet de variables : chol, thalach et oldpeak, nous tirons les conclusions suivantes :

1. L'équation du modèle de régression logistique nous montre que les variables chol et oldpeak influencent négativement la probabilité qu'une personne soit malade et la variable thalach influence positivement la probabilité qu'une personne soit malade.
2. Les variables chol et oldpeak ayant un rapport de côte inférieur à 1, repectivement $OR_{\text{chol}} = 0.993$ et $OR_{\text{oldpeak}} = 0.446$, plus leur valeur sera faible, plus la probabilité qu'une personne soit malade sera importante.
3. La variable thalach ayant un rapport de cote supérieur à 1, $OR_{\text{thalach}} = 1.036$, plus sa valeur sera importante plus la probabilité qu'une personne soit malade va augmenter.

En somme, si une personne à un rythme cardiaque maximal élevé, la probabilité qu'une personne soit malade augmente. Si une personne à un taux de cholestérol faible et un rythme cardiaque après l'effort faible (par rapport à un sujet sain), moins la probabilité qu'elle soit malade sera importante.

4.3 Analyse des clusters

L'analyse des clusters consiste à organiser les différentes variables de manière à former plusieurs clusters d'éléments en fonction de leur degré de similitudes.

4.3.1 Sous SAS

En SAS, la proc clusters va d'abord estimer le nombre de groupes qu'il faudra créer. Ci-dessous se trouve ce que retourne la proc cluster :

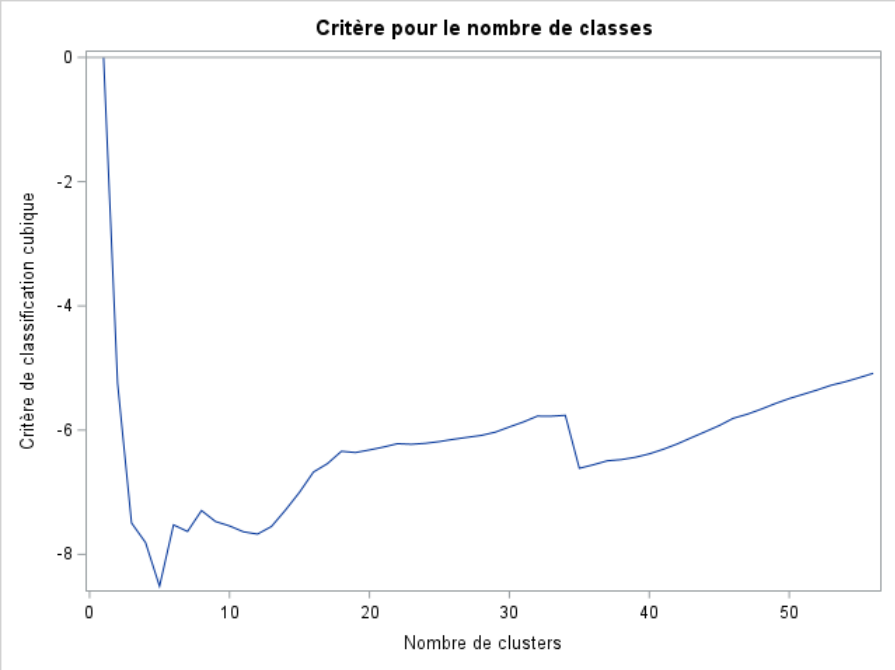


FIGURE 23 – Courbe permettant de définir le nombre de clusters pour la méthode

Ici, il semble évident que le nombre de clusters à sélectionner sera de 5. Au vu des résultats précédents, les variables age et trestbps seront écartées de cette méthode. Une fois que le nombre de clusters idéal a été identifié, il faut alors à l'aide de la proc tree créer ces clusters. A l'aide d'une proc means, il est alors possible de comparer chaque moyenne de clusters sachant que plus la moyenne d'un clusters est faible, moins il y a d'individu malade dans le cluster.

On obtient les moyennes suivantes :

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Moyenne	0.543	0.640	0.628	0.200	0.509

TABLE 9 – Moyenne d'individu malade pour chaque clusters

Les clusters ayant les moyennes les plus élevés sont les clusters 2 et 3. Le cluster 4 est le cluster dont la moyenne de malade est de loin la plus faible (0.2). Ces clusters seront donc intéressant à étudier en détail. Ci dessous se trouvent les boxplots des trois variables chol, thalach et oldpeak pour chaque cluster.

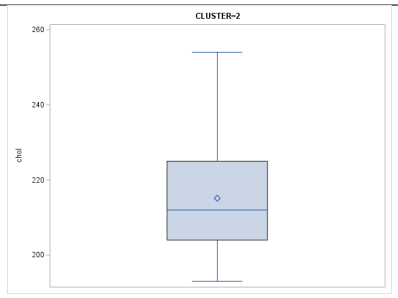
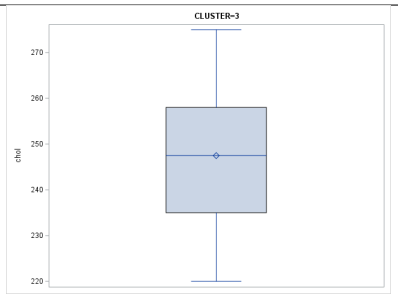
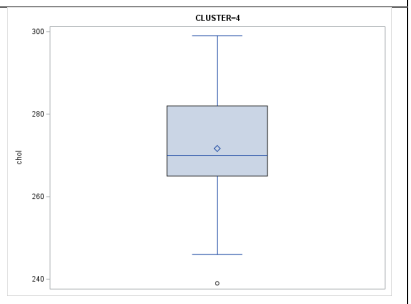
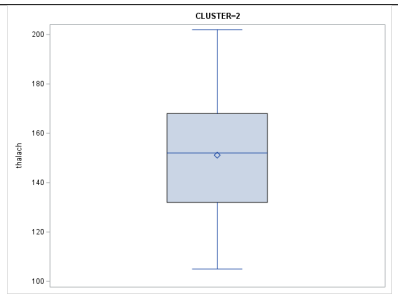
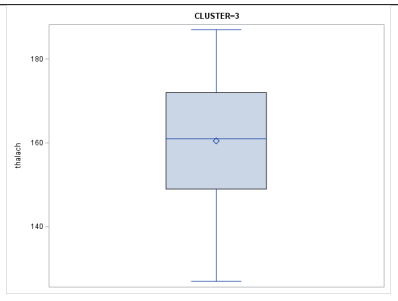
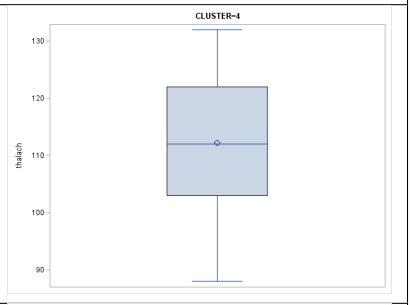
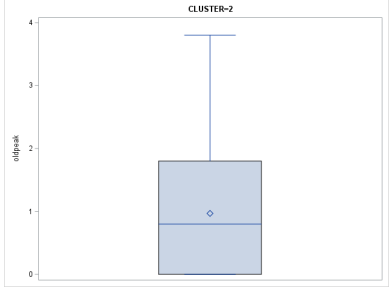
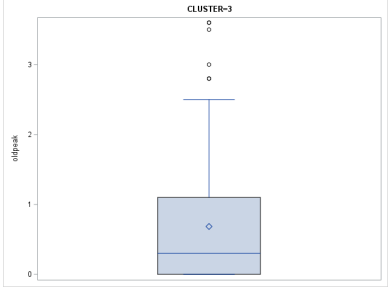
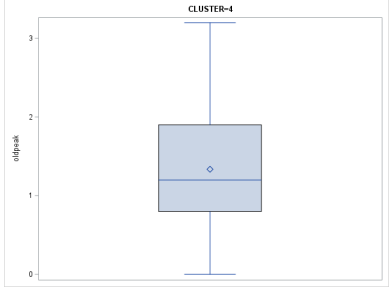
Var	Cluster 2	Cluster 3	Cluster 4
chol			
thalach			
oldpeak			

TABLE 10 – Boxplots de chaque cluster en fonction des variables chol, thalach et oldpeak

Des conclusions peuvent alors être émises sur l'impact des variables chol, thalach et oldpeak sur le risque de maladie cardiaque :

1. Plus le taux de cholestérol d'un individu est élevé, plus le risque qu'il soit atteint d'une maladie cardiaque est faible.
2. Plus le rythme cardiaque après un effort d'un individu est élevé, plus le risque qu'il soit atteint d'une maladie cardiaque est faible.
3. Plus le rythme cardiaque maximal d'un individu est élevé, plus le risque qu'il soit atteint d'une maladie cardiaque est fort.

4.3.2 Sous R

Contrairement à SAS, R nous a amené à travailler sur 4 clusters. Ceux-ci sont représenté ci-dessous :

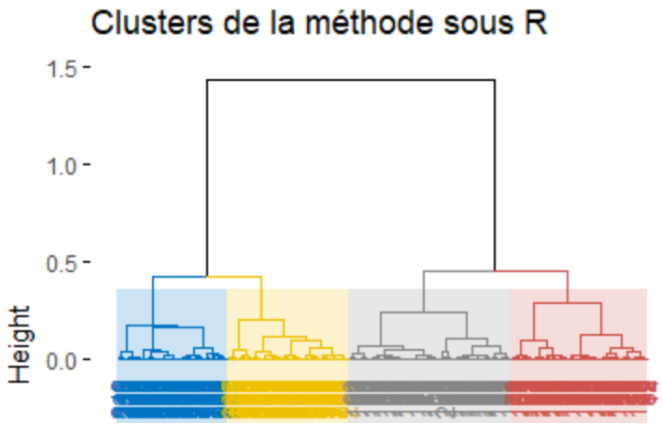


FIGURE 24 – Clusters qui vont être utilisé pour la méthode

La suite de la méthode consistera alors à trouver quels seront les clusters ayant la plus grandes populations de malades. Une fois que cela sera fait, il faudra alors trouver les variables les plus influentes pour ces clusters.

On trouve pour chaque cluster :

Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Malades	80.46 %	56.36 %	61.45 %	11.86 %

TABLE 11 – Pourcentage de malade dans la population de chaque clusters

Nous pouvons observer que les clusters 1 et 4 sont soit composés en grand partie d'individus atteints de maladie cardiaque (cluster 1), soit composés en grande partie d'individus en bonne santé (cluster 4). Pour les clusters 1 et 2, il semble que un peu plus de la moitié des individus soient malades.

Comme pour la partie SAS, seules les variables chol, thalach et oldpeak seront analysées.

Ci-dessous seront représenté les boxplots de chacune des variable pour chaque clusters. Ainsi, il sera possible d'établir des conclusions sur les variables qui semblent le plus impacter sur les clusters 1 et 4.

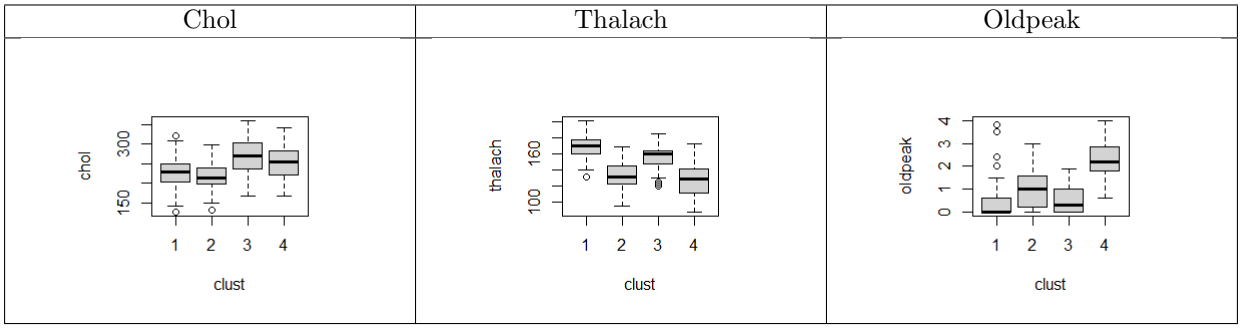


TABLE 12 – boxplots des variables pour chaque clusters

Les conclusions qui ressortent de cette table son les suivantes :

1. Pour la variable chol : Il semble que les individus du cluster 1 aient un taux de cholestérol moins élevés que les individus du cluster 4. Cela signifierait que plus le taux de cholestérol serait élevé, moins il y a de chance d'être atteint d'une maladie. Mais cela semble assez faible au vu des variations du taux de cholestérol entre ces deux boxplots et du taux de cholestérol dans les clusters 2 et 3.
2. Pour la variable thalach : Il est clairement observable que plus le rythme cardiaque maximal est élevé, plus il y a de chance que l'individu soit malade. En effet, cette variable est très élevée dans le cluster 1 et très faible dans le cluster 4.

3. Pour la variable oldpeak : Pour la variable oldpeak : Il est également observable que plus le rythme cardiaque après un effort est élevé, plus il y a de chance que l'individu soit en bonne santé. Les valeurs d'oldpeak sont les plus élevées pour le cluster 4 et les moins élevées pour le cluster 1.

4.4 Conclusion de l'analyse multivariée

Suite à l'analyse des composantes principales (ACP), il semblerait que les variables age et trestbps n'agissent pas sur les risques qu'un individu soit atteint d'une maladie cardiaque contrairement aux variables chol, thalach et oldpeak. Différentes observations ont été réalisées suites aux méthodes de régression logistique et des clusters :

— Plus thalach est élevé, plus il y a de chance que l'individu soit en bonne santé.

— Plus chol est élevé, plus il y a de chance que l'individu soit en bonne santé.

— Plus oldpeak est élevé, plus il y a de chance que l'individu soit atteint de maladie

La variation générée par la variable chol semble assez légère par rapport à celles de thalach et oldpeak.

5 Conclusion de l'étude

Au terme de cetté étude d'analyse visuelle, uni/bi/multi variée de données cardiaques, nous avons déterminé quelles variables allaient pouvoir expliquer le fait qu'une personne soit atteinte d'une malade cardiaque.

A travers des tests statistiques et des comparaison de distribution de valeurs nous avons validé ou non les hypothèses de normalité des variables. A travers une matrice de corrélation et des modèles de regression logistique nous avons découvert les relations qu'entretiennent les variables explicatives avec notre variable cible. En effet, nous avons conclu que seulement 3 des 5 variables explicatives initiales (chol, thalach, oldpeak) peuvent expliquer à elles seules la variable target. Enfin, à l'aide de l'analyse en composantes principales nous avons définitivement exclu les variables age et trestbps de l'étude. Nous avons ensuite montré, à l'aide d'un modèle de regression logistique que le taux de cholestérol et le rythme cardiaque après un effort sont inversement proprotionels à la probabilité qu'une personne soit malade i.e plus leurs valeurs sont faibles, plus la probabilité qu'une personne soit malade va augmenter. D'autre part, nous avons trouvé que le rythme cardiaque maximal est proportionel à la probabilité qu'une personne soit malade i.e plus sa valeur est importante plus la probabilité qu'une personne soit malade est importante. Une analyse par clusters nous à permis de solidifier nos conclusions en nous fournissant des résultats identiques à ceux obtenus dans la régression logistique multivariée.

Nous répondons donc à la problématique de la façon suivante : "Dans ce jeu de données, le fait qu'une personne soit malade sera expliqué par les données liées à son taux de cholestérol, son rythme cardiaque maximal et son rythme cardiaque après un effort."

Références

— Documentation SAS

— Documentation R

— Documentation Python