Analyse de Données

Rapport

Stave Icnel Dany OSIAS

2025-12-05

Table des matières

Liste des Tables

1 Travail personnel 1

Exercice 13.1 Données `Poids_naissance`

Chargeons les données `Poids_naissance` et donnons un aperçu de la matrice des données :

Table 1.1 – Aperçu des données `Poids_naissance`

ID	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FVT	BWT	LOW
85	19	182	2	0	0	0	1	0	2523	0
86	33	155	3	0	0	0	0	3	2551	0
87	20	105	1	1	0	0	0	1	2557	0
88	21	108	1	1	0	0	1	2	2594	0
89	18	107	1	1	0	0	1	0	2600	0
91	21	124	3	0	0	0	0	0	2622	0

Exercice 13.2 Conversion en même unité (`kg`)

Le poids des mères étant exprimé en `livres`, nous effectuons une transformation des données pour recoder cette variable en `kilogrammes` (1 livre = 0.45359237 kg).

Nous convertissons aussi le poids des bébés en kg pour avoir la même unité.

Table 1.2 – Aperçu après conversion du poids des mères

ID	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FVT	BWT	LOW
85	19	82.55381	2	0	0	0	1	0	2523	0
86	33	70.30682	3	0	0	0	0	3	2551	0
87	20	47.62720	1	1	0	0	0	1	2557	0
88	21	48.98798	1	1	0	0	1	2	2594	0
89	18	48.53438	1	1	0	0	1	0	2600	0
91	21	56.24545	3	0	0	0	0	0	2622	0

Table 1.3 – Aperçu après conversion du poids des bébés

ID	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FVT	BWT	LOW
85	19	82.55381	2	0	0	0	1	0	2.523	0
86	33	70.30682	3	0	0	0	0	3	2.551	0
87	20	47.62720	1	1	0	0	0	1	2.557	0
88	21	48.98798	1	1	0	0	1	2	2.594	0
89	18	48.53438	1	1	0	0	1	0	2.600	0
91	21	56.24545	3	0	0	0	0	0	2.622	0

Exercice 13.3 Tris à plat

Extrayons les variables.

On va d'abord classifier nos variables selon leurs types.

Ensuite, on fera des tris à plat que sur nos variables catégorielles et nos variables numériques discrètes.

Et pour nos variables numériques continues, on affichera les paramètres statistiques de base avec `summary`.

Table 1.4 – Tri à plat de la variable RACE

Modalités	Effectifs
1	96
2	26
3	67

Table 1.5 – Tri à plat de la variable SMOKE

Modalités	Effectifs
0	115
1	74

Table 1.6 – Tri à plat de la variable HT

Modalités	Effectifs
0	177
1	12

Table 1.7 – Tri à plat de la variable UI

Modalités	Effectifs
0	161
1	28

Table 1.8 – Tri à plat de la variable LOW

Modalités	Effectifs
0	130
1	59

Table 1.9 – Tri à plat de la variable PTL

Modalités	Effectifs
0	159
1	24
2	5
3	1

Table 1.10 – Tri à plat de la variable FVT

Modalités	Effectifs
0	100
1	47
2	30
3	7
4	4
6	1

Table 1.11 – Résumé statistique de la variable AGE

Statistique	Valeur
Min.	14.00
1st Qu.	19.00
Median	23.00
Mean	23.24
3rd Qu.	26.00
Max.	45.00

Table 1.12 – Résumé statistique de la variable LWT

Statistique	Valeur
Min.	36.29
1st Qu.	49.90
Median	54.88
Mean	58.88
3rd Qu.	63.50
Max.	113.40

Table 1.13 – Résumé statistique de la variable BWT

Statistique	Valeur
Min.	0.71
1st Qu.	2.41
Median	2.98
Mean	2.94
3rd Qu.	3.48
Max.	4.99

Exercice 14.1 Données `acteurs`

Créons le tableau de données `acteurs` :

Table 1.14 – Données `acteurs`

Mort.à	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.décès
93	66	211	Michel	Galabru	04-01-2016
53	25	58	André	Raimbourg	23-09-1970
72	48	98	Jean	Gabin	15-10-1976
68	37	140	Louis	De Funès	27-01-1983
68	31	74	Lino	Ventura	22-10-1987
53	32	81	Jacques	Villeret	28-01-2005

Exercice 14.2 Changement du nom de la 1ère colonne

Changeons le nom de la 1ère colonne par : Age.du.décès

Table 1.15 – Données `acteurs` après le changement de nom

Age.du.décès	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.décès
93	66	211	Michel	Galabru	04-01-2016
53	25	58	André	Raimbourg	23-09-1970
72	48	98	Jean	Gabin	15-10-1976
68	37	140	Louis	De Funès	27-01-1983
68	31	74	Lino	Ventura	22-10-1987
53	32	81	Jacques	Villeret	28-01-2005

Exercice 14.3 Extraction de la colonne `Prénom`

Extrayons la colonne `Prénom` :

Table 1.16 – Colonne `Prénom`

Prénom
Michel
André
Jean
Louis
Lino
Jacques

Exercice 14.4 Ordre croissant suivant l'âge du décès

Ordonnons le jeu de données `acteurs` par ordre croissant suivant l'âge du décès :

Table 1.17 – Données `acteurs` : tri croissant selon l'âge au décès

	Age.du.décès	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.décès
2	53	25	58	André	Raimbourg	23-09-1970
6	53	32	81	Jacques	Villeret	28-01-2005
4	68	37	140	Louis	De Funès	27-01-1983

	Age.du.décès	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.décès
5	68	31	74	Lino	Ventura	22-10-1987
3	72	48	98	Jean	Gabin	15-10-1976
1	93	66	211	Michel	Galabru	04-01-2016

Exercice 15.1 Données `fromages`

Construisons une data frame w constituée du jeu de données `fromages` avec les noms des colonnes :

Table 1.18 – Aperçu des données `fromages`

Y	X1	X2	X3
12.3	4.543	3.135	0.86
20.9	5.159	5.043	1.53
39.0	5.366	5.438	1.57
47.9	5.759	7.496	1.81
5.6	4.663	3.807	0.99
25.9	5.697	7.601	1.09

Exercice 15.2 Utilisation de `attach`

Associons (attachons) les noms aux colonnes respectives. Vérifions en tapant X1 :

Table 1.19 – Aperçu de la colonne `X1` après `attach`

X1
4.543
5.159
5.366
5.759
4.663
5.697

Exercice 15.3 Caractéristiques de w

Affichons les caractéristiques de w avec `names` et `summary` :

Table 1.20 – Noms des variables de w

Variable
Y
X1
X2
X3

Table 1.21 – Structure de w

Variable	Type	Valeurs
Y	num	12.3 20.9 39 47.9 5.6 25.9 37.3 21.9 18.1 21 ...
X1	num	4.54 5.16 5.37 5.76 4.66 ...
X2	num	3.13 5.04 5.44 7.5 3.81 ...
X3	num	0.86 1.53 1.57 1.81 0.99 1.09 1.29 1.78 1.29 1.58 ...

Exercice 15.4 Résumé statistique des variables

Donnons les paramètres statistiques élémentaires pour les variables Y , X1, X2 et X3.

Table 1.22 – Résumé statistique

Statistique	Valeur
Min.	0.70
1st Qu.	13.55
Median	20.95
Mean	24.53
3rd Qu.	36.70
Max.	57.20

Table 1.23 – Résumé statistique

Statistique	Valeur
Min.	4.48
1st Qu.	5.24
Median	5.42
Mean	5.50
3rd Qu.	5.88
Max.	6.46

Table 1.24 – Résumé statistique

Statistique	Valeur
Min.	3.00
1st Qu.	3.98
Median	5.33
Mean	5.94
3rd Qu.	7.57
Max.	10.20

Table 1.25 – Résumé statistique

Statistique	Valeur
Min.	0.86
1st Qu.	1.25
Median	1.45
Mean	1.44
3rd Qu.	1.67
Max.	2.01

Exercice 15.5 Commande `pairs`

Exécutons la commande `pairs(w)`. Cela nous renvoie tous les nuages de points de toutes les variables par paire :

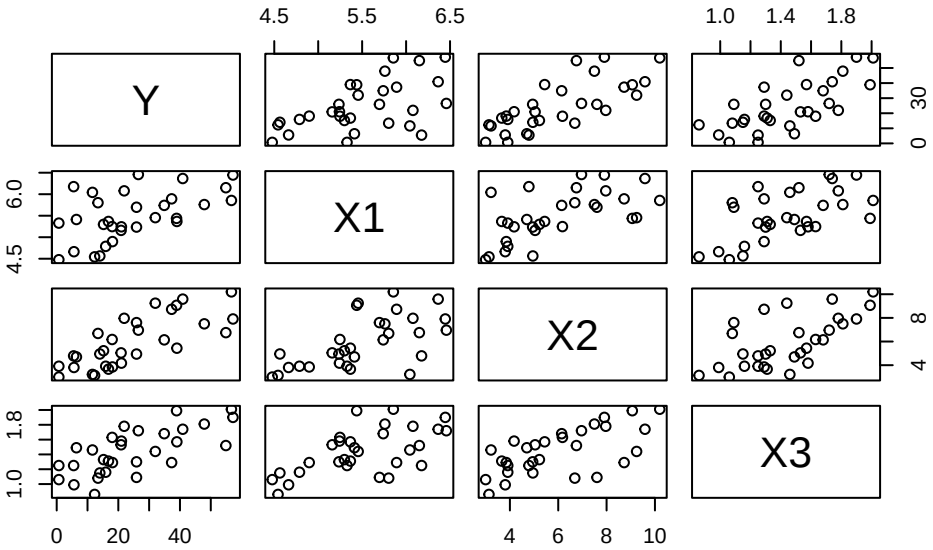


Figure 1.1 – Nuages de points par paire

Exercice 15.6 Jeu de données `ww`

Construisons maintenant une nouvelle data.frame ww des individus vérifiant X1 > 5.1 et X3 < 1.77 et affichons les premieres lignes :

Table 1.26 – Aperçu des données `ww`

	Y	X1	X2	X3
2	20.9	5.159	5.043	1.53
3	39.0	5.366	5.438	1.57
6	25.9	5.697	7.601	1.09

	Y	X1	X2	X3
7	37.3	5.892	8.726	1.29
10	21.0	5.242	4.174	1.58
11	34.9	5.740	6.142	1.68

Exercice 15.7 Caractéristiques de ww

Affichons les caractéristiques de ww :
Dimensions de ww : 19 4

Table 1.27 – Structure des données

Variable	Type	Valeurs
Y	num	20.9 39 25.9 37.3 21 34.9 25.9 54.9 40.9 6.4 ...
X1	num	5.16 5.37 5.7 5.89 5.24 ...
X2	num	5.04 5.44 7.6 8.73 4.17 ...
X3	num	1.53 1.57 1.09 1.29 1.58 1.68 1.3 1.52 1.74 1.49 ...

Exercice 15.8 Résumé statistique des variables à partir de ww

A partir de ww, donnons les paramètres statistiques élémentaires pour les variables Y , X1, X2 et X3 :

Table 1.28 – Résumé statistique de : Y

Statistique	Valeur
Min.	0.70
1st Qu.	14.30
Median	21.00
Mean	23.52
3rd Qu.	33.45
Max.	54.90

Table 1.29 – Résumé statistique de : X1

Statistique	Valeur
Min.	5.16
1st Qu.	5.31
Median	5.46
Mean	5.65
3rd Qu.	5.97
Max.	6.46

Table 1.30 – Résumé statistique de : X2

Statistique	Valeur
Min.	3.22
1st Qu.	4.74
Median	5.44
Mean	5.95
3rd Qu.	6.86
Max.	9.59

Table 1.31 – Résumé statistique de : X3

Statistique	Valeur
Min.	1.08
1st Qu.	1.29
Median	1.46
Mean	1.43
3rd Qu.	1.58
Max.	1.74

Exercice 16.1 Données `airquality`

Chargeons les données `airquality` :

Table 1.32 – Aperçu des données `airquality`

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

- Ce jeu de données provient de mesures de la qualité de l'air à New York entre le mois de Mai et le mois de Septembre 1973.

Exercice 16.2 Noms des variables de `airquality`

Affichons les noms des variables :

```
Noms des variables de `airquality`:  Ozone Solar.R Wind Temp Month Day
```

Exercice 16.3 Nombre de lignes et de colonnes

Affichons le nombre de lignes et de colonnes :

```
Nombre de lignes:  153
```

```
Nombre de colonnes:  6
```

Exercice 16.4 Résumé statistique de `airquality`

Calculons les paramètres statistiques de base à l'aide de la commande `summary` :

Table 1.33 – Résumé statistique de Ozone

Statistique	Valeur
Min.	1.00
1st Qu.	18.00
Median	31.50
Mean	42.13
3rd Qu.	63.25
Max.	168.00
NA's	37.00

Table 1.34 – Résumé statistique de Solar.R

Statistique	Valeur
Min.	7.00
1st Qu.	115.75
Median	205.00
Mean	185.93
3rd Qu.	258.75
Max.	334.00
NA's	7.00

Table 1.35 – Résumé statistique de Wind

Statistique	Valeur
Min.	1.70
1st Qu.	7.40
Median	9.70
Mean	9.96
3rd Qu.	11.50
Max.	20.70

Table 1.36 – Résumé statistique de Temp

Statistique	Valeur
Min.	56.00
1st Qu.	72.00
Median	79.00
Mean	77.88
3rd Qu.	85.00

Max.	97.00

Table 1.37 – Résumé statistique de Month

Statistique	Valeur
Min.	5.00
1st Qu.	6.00
Median	7.00
Mean	6.99
3rd Qu.	8.00
Max.	9.00

Table 1.38 – Résumé statistique de Day

Statistique	Valeur
Min.	1.0
1st Qu.	8.0
Median	16.0
Mean	15.8
3rd Qu.	23.0
Max.	31.0

Exercice 16.5 Boîtes à moustaches par mois de `Ozone`

Représentons les boîtes à moustaches de la variable `Ozone` pour chaque mois avec la commande `plot` :

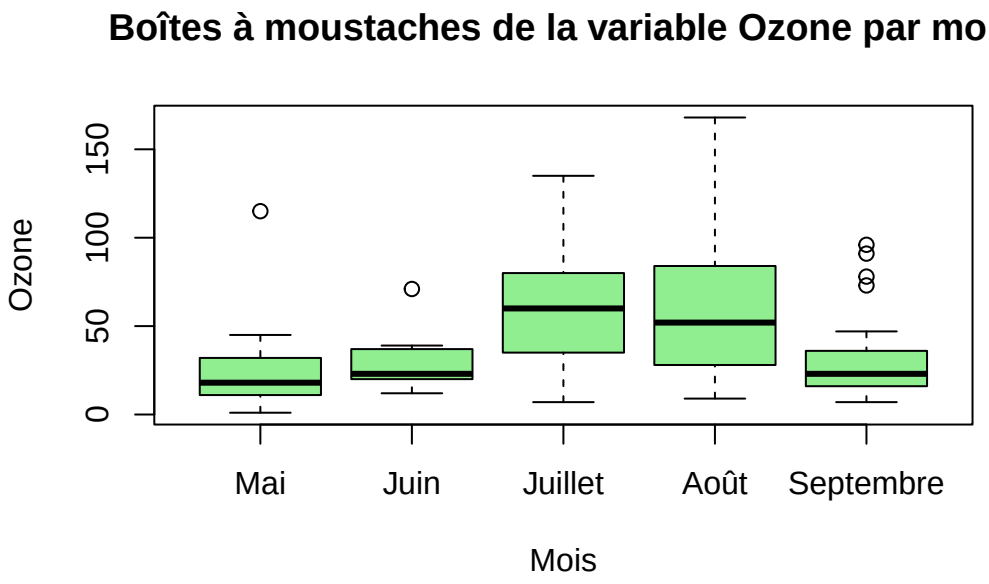


Figure 1.2 – Boîte à moustaches - Ozone par mois

- Remarquons que, pour les mois de Juillet et d'Août, le niveau d'ozone a une plus grande variance et pas de valeurs extrêmes.
- La distribution d'ozone a le plus de valeurs extrêmes pour le mois de septembre.
- Le niveau d'ozone varie très peu pour le mois de juin.

Exercice 16.6 Création de la variable qualitative `saison`

Créons une variable qualitative `saison` qui vaut `printemps` quand le mois est 5 (Mai), `été` quand les mois sont 6, 7 et 8 (Juin, Juillet, Août), et `automne` quand le mois est 9 (Septembre).

Table 1.39 – Aperçu des données `airquality` avec la variable `saison`

Ozone	Solar.R	Wind	Temp	Month	Day	Saison
41	190	7.4	67	Mai	1	Printemps
36	118	8.0	72	Mai	2	Printemps
12	149	12.6	74	Mai	3	Printemps
18	313	11.5	62	Mai	4	Printemps
NA	NA	14.3	56	Mai	5	Printemps
28	NA	14.9	66	Mai	6	Printemps

Exercice 16.7 Niveau d'ozone selon la température

Obtenons le nuage de points du niveau d'ozone selon la température :

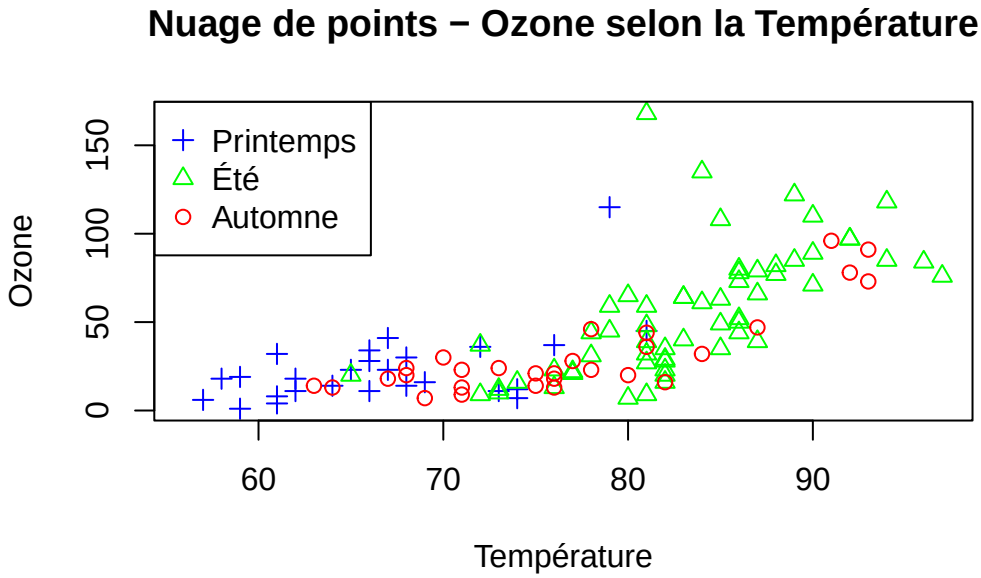


Figure 1.3 – Nuage de points - Ozone selon la Température

Exercice 17.1 Simulation d'une distribution normale

Simulons 100 valeurs $e_1, ..., e_{100}$ d'une var suivant la loi normale $N(0, 5^2)$:

Table 1.40 – Aperçu des e_i

x
-1.975309
-2.690974
-3.769857
-9.253822
-7.450896
-2.995864

Exercice 17.2 Création des y_i

Pour tout $i \in \{1, \dots, 100\}$, on pose $y_i = 1.7 + 2.1i + e_i$. On obtient au final le vecteur y :

Table 1.41 – Aperçu des y_i

x
1.8246912
3.2090264
4.2301429
0.8461779
4.7491042

x
11.3041357

Exercice 17.3 Nuage de points (i, y_i)

Représentons le nuage de points (i, y_i) pour $i \in \{1, \dots, 100\}$.

Sur ce même graphique, ajoutons en rouge la droite qui ajuste au mieux ce nuage de points, autrement dit la droite de régression.

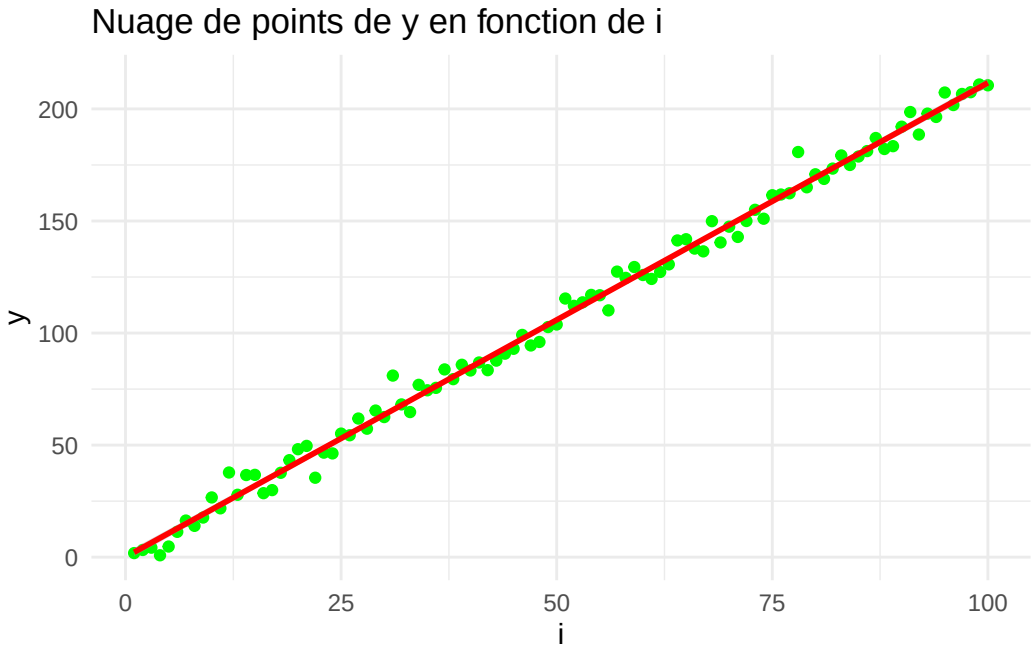


Figure 1.4 – Nuage de points de y_i en fonction de i avec droite de régression

- Les y_i étant déjà définis linéairement par rapport à i, la droite de régression linéaire simple est bien la droite qui ajuste au mieux le nuage de points.

Exercice 18.1 Tableau de contingence : couleur des yeux et couleur de cheveux

Créons le tableau de contingence croisant la couleur des yeux et la couleur de cheveux pour 592 femmes :

Table 1.42 – Tableau de contingence : couleur des yeux et couleur de cheveux

	brun	chatin	roux	blond
marron	68	119	26	7
noisette	15	54	14	10
vert	5	29	14	16
bleu	20	84	17	94

Exercice 18.2 Tableau de contingence des fréquences

Calculons la matrice des fréquences (arrondir au 100ème près) :

```
Total des effectifs:  592
```

Table 1.43 – Tableau de contingence des fréquences

	brun	chatin	roux	blond
marron	0.11	0.20	0.04	0.01
noisette	0.03	0.09	0.02	0.02
vert	0.01	0.05	0.02	0.03
bleu	0.03	0.14	0.03	0.16

Exercice 18.3 Lois marginales

Donnons les lois marginales (c pour le vecteur colonne et r pour le vecteur ligne) :

Table 1.44 – Loi marginale r par ligne (couleur des yeux)

x
0.36
0.16
0.11
0.36

Table 1.45 – Loi marginale c par colonne (couleur de cheveux)

x
0.18
0.48
0.11
0.22

Exercice 18.4 Matrice des profils-lignes L

Utilisons la commande `sweep` pour donner la matrice des profils lignes L (distributions conditionnelles en ligne) :

Table 1.46 – Profils-lignes (couleur des yeux)

	brun	chatin	roux	blond
marron	0.31	0.56	0.11	0.03
noisette	0.19	0.56	0.12	0.12
vert	0.09	0.45	0.18	0.27
bleu	0.08	0.39	0.08	0.44

Exercice 18.5 Matrice des profils-colonnes C

Utilisons la commande `sweep` pour donner la matrice des profils colonnes C (distributions conditionnelles en colonne) :

Table 1.47 – Profils-colonnes (couleur des cheveux)

	brun	chatin	roux	blond
marron	0.61	0.42	0.36	0.05
noisette	0.17	0.19	0.18	0.09
vert	0.06	0.10	0.18	0.14
bleu	0.17	0.29	0.27	0.73

Exercice 18.6 Distance de χ^2 manuelle

Calculons manuellement la distance de χ^2 entre les profils lignes :

Table 1.48 – Distance de χ^2 manuelle

	marron	noisette	vert	bleu
marron	0.000	0.118	0.600	1.126
noisette	0.118	0.000	0.216	0.607
vert	0.600	0.216	0.000	0.230
bleu	1.126	0.607	0.230	0.000

Exercice 18.7 Matrice des taux de liaison

Donnons la matrice des taux de liaison (arrondir au 100ème près) :

Table 1.49 – Matrice des taux de liaison

	brun	chatin	roux	blond
marron	0.18	0.07	0.00	-0.25
noisette	0.01	0.05	0.02	-0.08
vert	-0.07	-0.01	0.07	0.04
bleu	-0.14	-0.08	-0.05	0.29

Exercice 18.8 Test de χ^2

Faisons le test de χ^2 permettant de juger de la liaison entre la couleur des yeux et la couleur des cheveux :

```
    Pearson's Chi-squared test

data:  data
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

- La p-valeur résultant du test de χ^2 est inférieure a 0.05. On **rejette** donc l'hypothese d'indépendance.
- La couleur des yeux et la couleur des cheveux sont statistiquement significativement dépendantes.

2 Travail personnel 2

Exercice 19.1. Tableau de contingence : Niveau de Diplôme et Tranche d'âge

Ecrivons un tableau comportant les données :

Table 2.1 – Tableau de contingence des effectifs : Niveau de Diplôme et Tranche d'âge

	BEPC	BAC	Licence	Total
Plus de 50 ans	15	12	3	30
Entre 30 et 50 ans	10	18	4	32
Moins de 30 ans	15	5	8	28
Total	40	35	15	90

Exercice 19.2. Tableau de contingence des fréquences

Donnons le tableau de contingence des fréquences :

Table 2.2 – Tableau de contingence des fréquences : Niveau de Diplôme et Tranche d'âge

	BEPC	BAC	Licence	Fréq. relat. colonne
Plus de 50 ans	0.167	0.133	0.033	0.333
Entre 30 et 50 ans	0.111	0.200	0.044	0.356
Moins de 30 ans	0.167	0.056	0.089	0.311
Fréq. relat. ligne	0.444	0.389	0.167	1.000

Exercice 19.3. Matrices des profils-lignes et profils-colonnes

Donnons la matrice des profils lignes et celle des profils colonnes :

Table 2.3 – Matrice des profils-lignes

	BEPC	BAC	Licence	Total
Plus de 50 ans	0.500	0.400	0.100	1
Entre 30 et 50 ans	0.312	0.562	0.125	1
Moins de 30 ans	0.536	0.179	0.286	1

Table 2.4 – Matrice des profils-colonnes

	BEPC	BAC	Licence
Plus de 50 ans	0.375	0.343	0.200
Entre 30 et 50 ans	0.250	0.514	0.267
Moins de 30 ans	0.375	0.143	0.533
Total	1.000	1.000	1.000

Exercice 19.4. Test d'indépendance de χ^2

Effectuons le test de χ^2 d'indépendance des deux caractères statistiques :

```
    Pearson's Chi-squared test

data:  tab
X-squared = 11.175, df = 4, p-value = 0.02466
```

Exercice 19.5. Indépendance pour un seuil $\alpha = 0.05$

Les deux caractères statistiques sont-ils indépendants pour un seuil $\alpha = 0.05$?

- La p-value étant inférieure à 0.05, on **rejette** l'hypothèse nulle d'indépendance.
- Il existe une relation statistiquement significative entre l'âge et le diplôme.

Exercice 20.1. Tableau de contingence : `Etat matrimonial` et `Couleur des yeux`

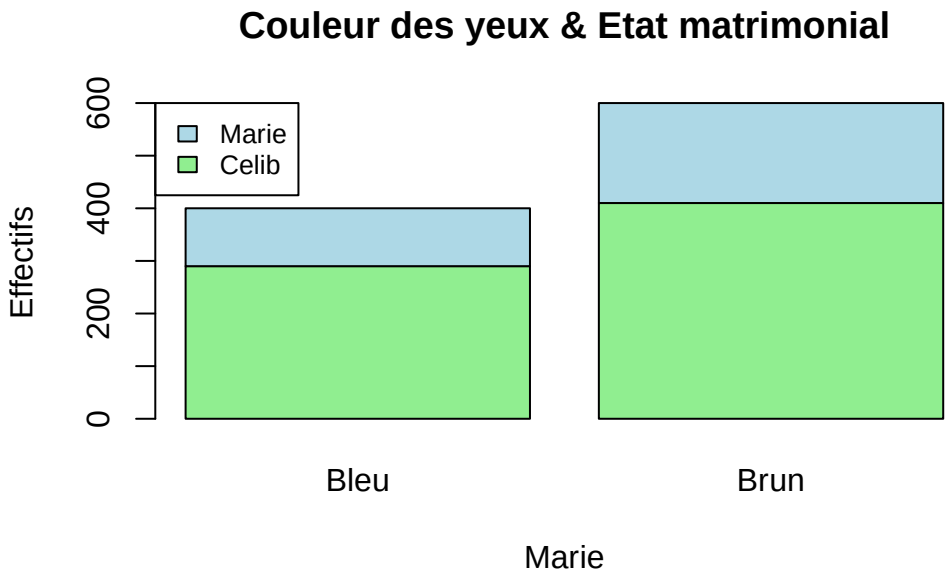
Créons une variable `tableau` pour le tableau de contingence des effectifs :

Table 2.5 – Tableau de contingence : `Etat matrimonial` et `Couleur des yeux`

	Bleu	Brun
Celib	290	410
Marie	110	190

Exercice 20.2. Représentation graphique du tableau

Représentons graphiquement ce tableau de contingence des effectifs avec `barplot` :



Exercice 20.3. Commandes `margin.table` et `prop.table`

Exécutons les commandes `margin.table(tableau)`, `margin.table(tableau,1)`, `margin.table(tableau,2)` et `prop.table(tableau)` :

```
margin.table(tableau):  1000
```

Table 2.6 – Distribution marginale de la première variable `Etat matrimonial`

Var1	Freq
Celib	700
Marie	300

Table 2.7 – Distribution marginale de la deuxième variable `Couleur des yeux`

Var1	Freq
Bleu	400
Brun	600

Table 2.8 – Tableau de contingence des fréquences

	Bleu	Brun
Celib	0.29	0.41
Marie	0.11	0.19

- `margin.table(tableau)` renvoie l'effectif total de l'échantillon.
- `margin.table(tableau,1)` renvoie la distribution marginale de la premiere variable `Etat matrimonial`.
- `margin.table(tableau,2)` renvoie la distribution marginale de la deuxieme variable `Couleur des yeux`.
- `prop.table(tableau)` renvoie le tableau de contingence des fréquences.

Exercice 20.4. Tableau `tab0` et résumé statistique

a) Créons le tableau `tab0` :

Table 2.9 – `tab0`

	Bleu	Brun
Celib	280	420
Marie	120	180

- `tab0`, ainsi défini, est le **tableau des effectifs théoriques** sous l'hypothèse d'indépendance.

b) Exécutons la commande `summary` sur nos 2 tableaux :

```
Number of cases in table: 1000
Number of factors: 2
Test for independence of all factors:
    Chisq = 1.9841, df = 1, p-value = 0.159

Number of cases in table: 1000
Number of factors: 2
Test for independence of all factors:
    Chisq = 1.154e-29, df = 1, p-value = 1
```

- Les p-valeurs des tests de χ^2 sur les 2 tableaux sont toutes deux supérieures à 0.05. Dans ce cas, on **conserve** donc l'hypothèse nulle d'indépendance des deux variables `Etat matrimonial` et `Couleur des yeux`.
- `tab0` a, dès le départ, été construit comme le tableau des effectifs théoriques sous l'hypothèse d'indépendance. Ce n'est donc pas surprenant que la p-valeur, résultant du test de χ^2, est égale à 1.

Exercice 20.5. `tableau2` et test de χ^2

Créons un tableau dans lequel tous les individus aux yeux `bleus` sont `mariés` et tous les autres sont `célibataires`, et effectuons le test de χ^2 :

Table 2.10 – `tableau2`

	Bleu	Brun
Celib	0	600
Marie	400	0

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  tableau2
X-squared = 995.84, df = 1, p-value < 2.2e-16
```

- Dans ce cas, la p-valeur résultant du test de χ^2 est bien inférieure à 0.05. Par conséquent, on **rejette** l'hypothèse nulle d'indépendance.
- Pour ce tableau de contingence, on dira qu'il existe une relation statistiquement significative entre les deux variables.

Exercice 20.6. Test de χ^2 sur `HairEyeColor`, `Titanic` et `UCBAdmissions`

b) Appliquons le test de χ^2 à quelques échantillons statistiques de R, par exemple `HairEyeColor`, `Titanic` et `UCBAdmissions`.

Table 2.11 – Tableau de contingence : `Hair` & `Eye`

	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

```
    Pearson's Chi-squared test

data:  tab_hair_eye
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

Table 2.12 – Tableau de contingence : `Class` & `Survived`

	No	Yes
1st	122	203
2nd	167	118
3rd	528	178
Crew	673	212

```
    Pearson's Chi-squared test

data:  tab_class_surv
X-squared = 190.4, df = 3, p-value < 2.2e-16
```

Table 2.13 – Tableau de contingence : `Admit` & `Dept`

	A	B	C	D	E	F
Admitted	601	370	322	269	147	46
Rejected	332	215	596	523	437	668

```
    Pearson's Chi-squared test

data:  tab_admit_dept
X-squared = 778.91, df = 5, p-value < 2.2e-16
```

Commentaire

- Toutes les p-valeurs obtenues sont inférieures à 0.05 ; on **rejette** donc l'hypothèse d'indépendance des paires de variables qu'on a choisies pour chacun des 3 échantillons statistiques : `HairEyeColor`, `Titanic` et `UCBAdmissions`.
- Dans le cas de `HairEyeColor`, on peut donc dire que la **couleur des cheveux** et la **couleur des yeux** sont signifcativement corrélées.
- Dans le cas de `Titanic`, on peut dire que la **classe** et la **survie** sont signifcativement corrélées.
- Dans le cas de `UCBAdmissions`, on peut dire que l'**admission** et le **département** sont signifcativement corrélées.

Exercice 21.2. Donnée `cars`

Dans cet exercice, nous allons non seulement manipuler les commandes de bases permettant d'effectuer une régression linéaire sous R, mais également contrôler les résultats obtenus, sélectionner les modèles et effectuer des représentations graphiques.

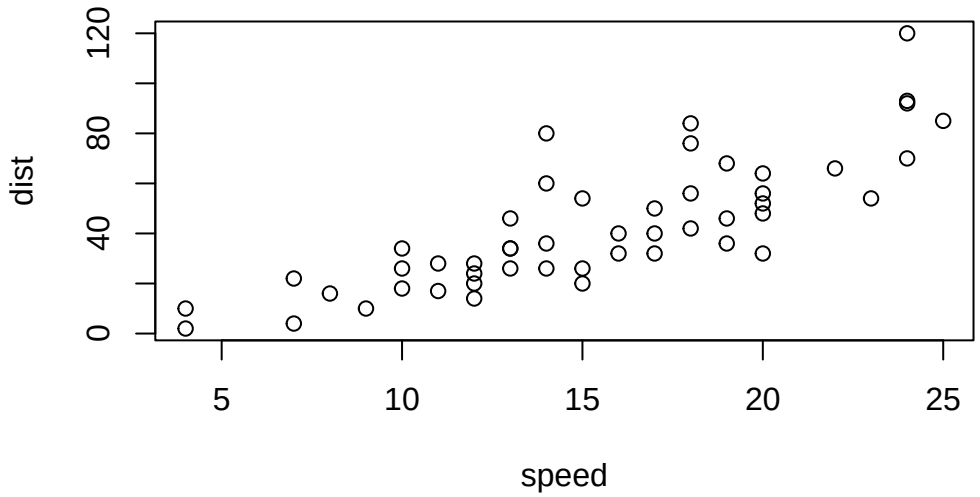
On va utiliser un jeu de données simple déjà implanté dans R : `cars`.

Table 2.14 – Aperçu des données `cars`

speed	dist
4	2
4	10
7	4
7	22
8	16
9	10

```
Noms de variables du jeu de données `cars`:  speed dist

Dimensions du jeu de données `cars`:  50 2
```

Puisque le jeu de données `cars` contient que 2 variables, on peut construire le nuage de points avec la fonction `plot` :



Ici, on construit un modèle de régression linéaire simple et on affichera ses attributs :

```
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"
```

Comparons `summary` et `anova` :

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

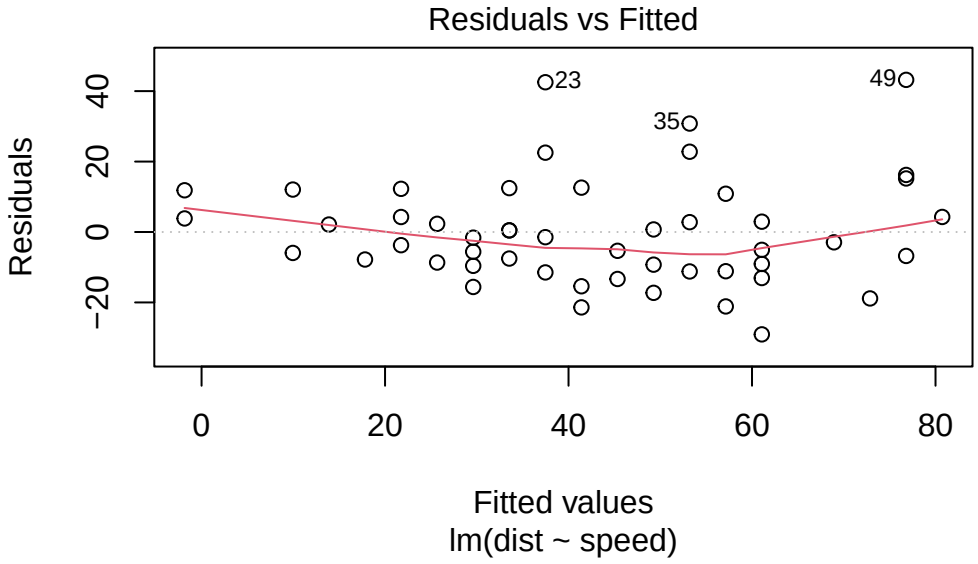
Commentaire

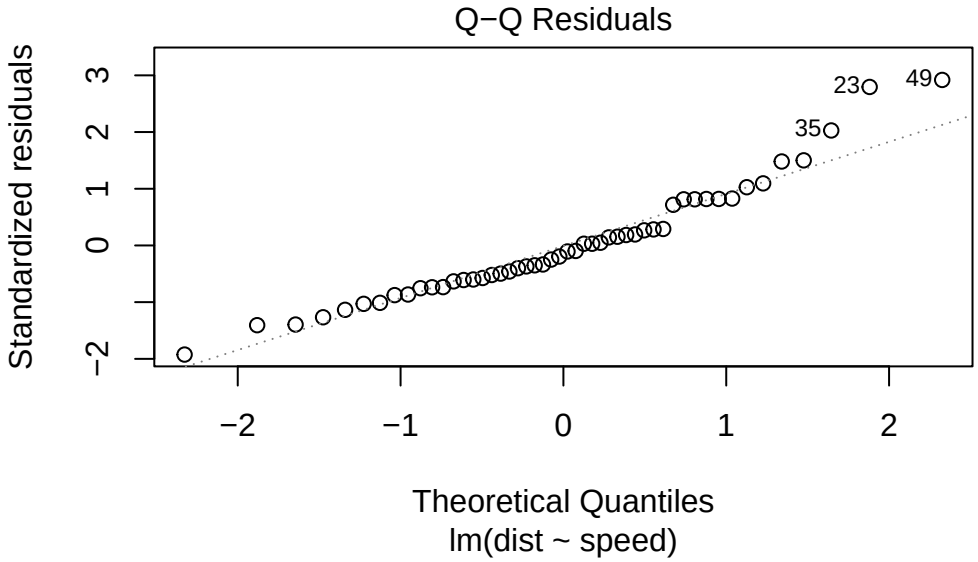
- `summary(reg)` fournit une **interprétation paramétrique** du modèle :
  - Estimation des coefficients (intercept et pente).
  - Test t de significativité des coefficients.
- `anova(reg)` fournit une **décomposition de la variance** :
  - Variance expliquée par `speed`.
  - Variance résiduelle.

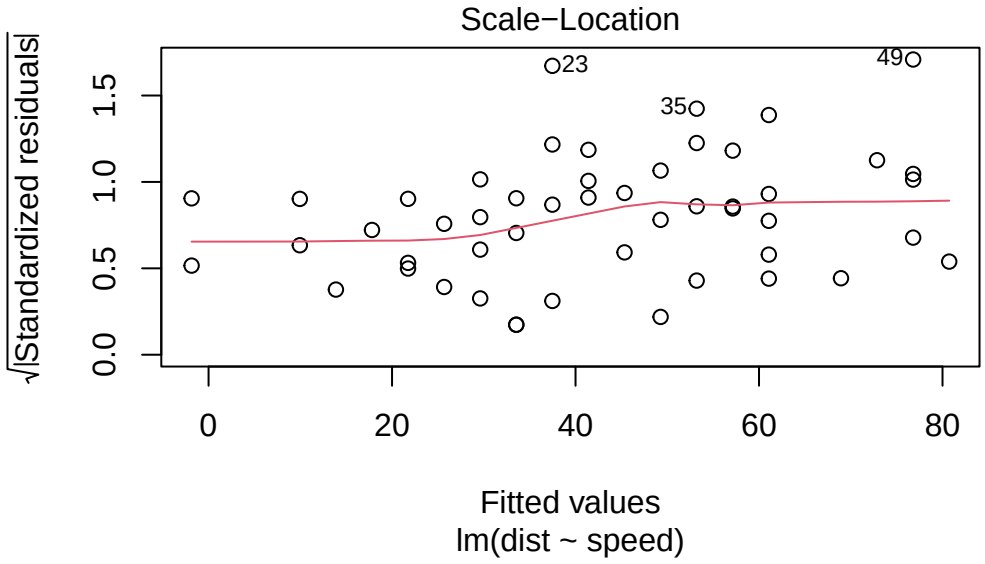
En gros, les deux méthodes testent l'effet de la variable explicative **speed** sur la variable réponse **dist**.

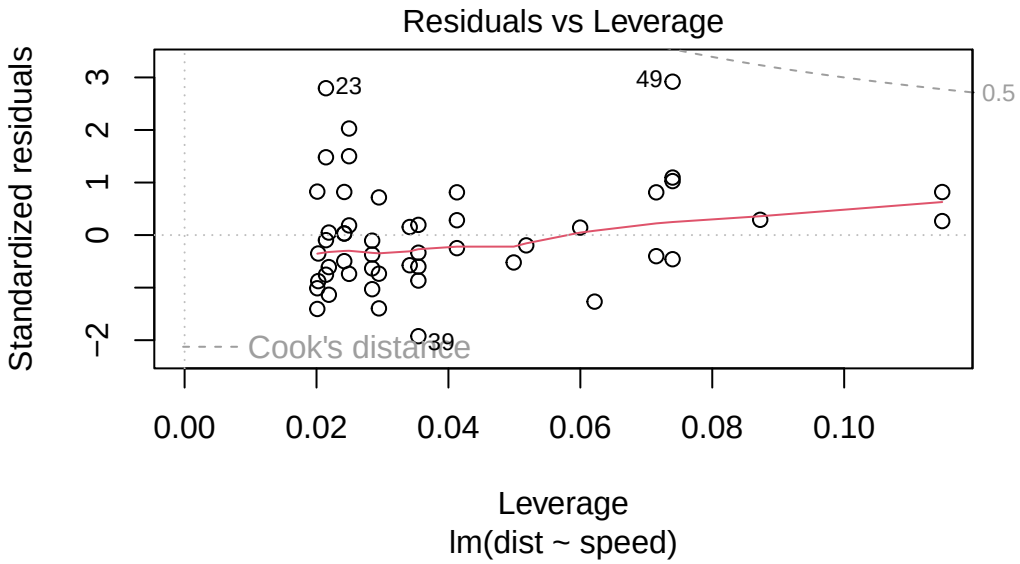
- La statistique de test est la même :
  - **F = 89.57**
  - **p-value = 1.49e-12**
- La conclusion est identique : l'effet de `speed` sur `dist` est **hautement significatif**.

La commande `plot(reg)` nous fournit 4 graphes mais on s'intéressera a deux en particulier :


Residuals vs Fitted
Residuals
Fitted values
lm(dist ~ speed)
23
35
49
−20
0
20
40
60
80


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)
23
35
49
−2
−1
0
1
2
3

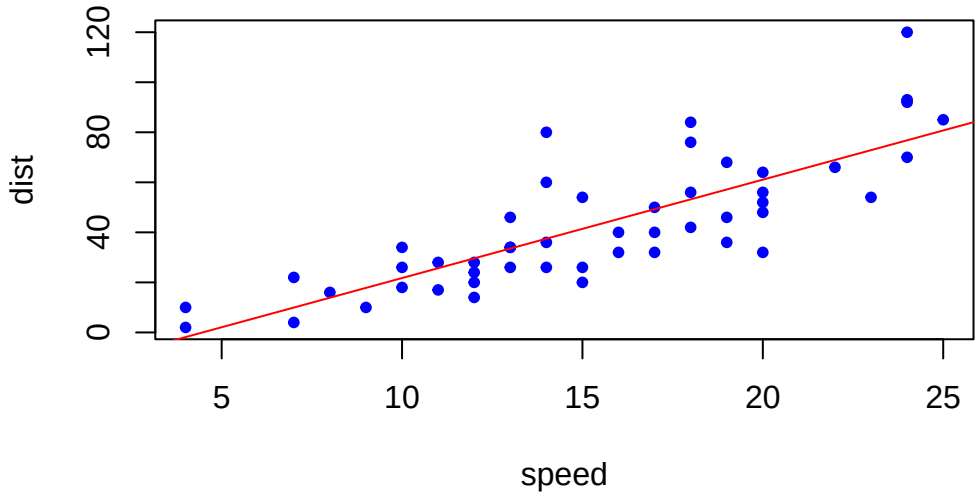




- Le graphique QQ-plot permet d'évaluer l'hypothèse de normalité des résidus du modèle.
    - Lorsque les points sont globalement alignés le long de la première bissectrice, cela indique que la distribution empirique des résidus est proche d'une loi normale.
    - Dans notre cas, la majorité des points suit correctement cette droite de référence. On observe toutefois de légers écarts dans les extrémités, traduisant la présence de quelques valeurs atypiques dans les queues de distribution.

- Le graphique de `Cook's D` permet de repérer les points ń influents ż, c'est-à-dire ceux pour qui la régression linéaire est mal (ou pas) adaptée, parce qu'ils se situent trop loin de la droite de régression. Ces points sont repérés par de grandes valeurs du D de Cook.

a) On peut tracer désormais la droite de régression :



- On peut voir que cette droite ajute bien nos données `cars`.

Pour la prévision, on a besoin de la commande `predict` :

```
La valeur prédite pour une vitesse de 20 est de:  61.06908
```

b) Donnons un intervalle de confiance et un intervalle de prédiction pour cette valeur avec les options `confidence` puis `prediction` :

```
       fit      lwr      upr
1 61.06908 55.24729 66.89088

       fit      lwr      upr
1 61.06908 29.60309 92.53507
```

- Remarquons que l'intervalle de confiance est beaucoup plus étroit que l'intervalle de prédiction.

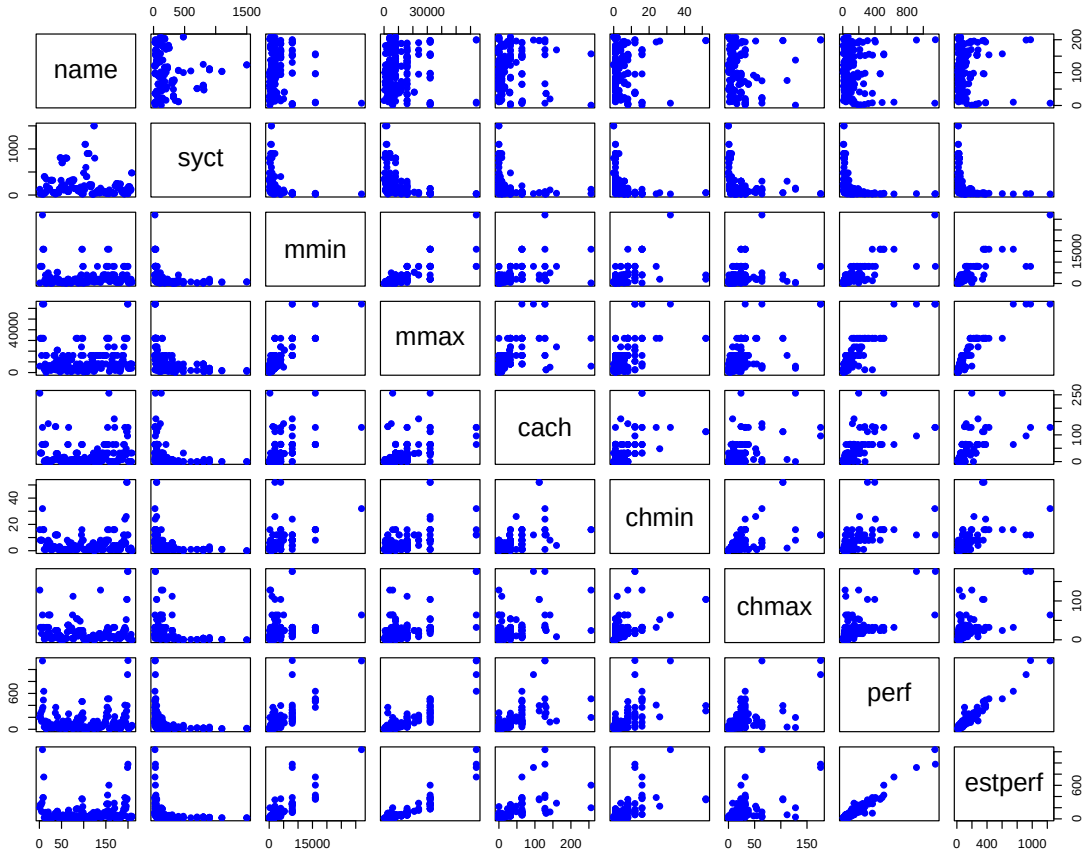
c) Pour le jeu de données `cars`, la sélection de variables n'est pas possible car il ne contient que deux variables : `speed` et `dist`.

Chargeons le jeu de données `cpus` du package `MASS` :

Table 2.15 – Aperçu des données `cpus`

name	syct	mmin	mmax	cach	chmin	chmax	perf	estperf
ADVISOR 32/60	125	256	6000	256	16	128	198	199
AMDAHL 470V/7	29	8000	32000	32	8	32	269	253
AMDAHL 470/7A	29	8000	32000	32	8	32	220	253
AMDAHL 470V/7B	29	8000	32000	32	8	32	172	253
AMDAHL 470V/7C	29	8000	16000	32	8	16	132	132
AMDAHL 470V/8	26	8000	32000	64	8	32	318	290

d) Pour avoir une idée globale du comportement des variables les unes par rapport aux autres, on utilise la fonction `pairs` :



- On peut remarquer que certaines variables comme `perf` et `estperf` sont fortement corrélées positivement.
- Pour les autres variables, les relations ne sont pas du tout linéaires.

Effectuons la régression de la variable `perf` contre toutes les autres variables quantitatives :

```
Call:
lm(formula = perf ~ ., data = cpus_quant)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-160.572  -15.224   -2.224    7.556  234.589

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.9069391  6.7801517   1.019   0.3096
syct        -0.0134520  0.0125081  -1.075   0.2835
mmin         0.0017772  0.0015114   1.176   0.2410
mmax        -0.0006548  0.0005910  -1.108   0.2692
cach         0.1740674  0.0990531   1.757   0.0804 .
chmin       -0.1072525  0.5786821  -0.185   0.8531
chmax        0.3479115  0.1657820   2.099   0.0371 *
estperf      0.9447315  0.0608743  15.519   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.56 on 201 degrees of freedom
Multiple R-squared:  0.9385,    Adjusted R-squared:  0.9364
F-statistic: 438.4 on 7 and 201 DF,  p-value: < 2.2e-16
```

Commentaire

- Le modèle est **globalement très significatif** (statistique de Fisher = 438.4, p-value < 2.2e-16).
- `estperf` est de loin la variable la plus explicative (p-value < 2.2e-16)
    - Cela confirme que la performance estimée est un excellent prédicteur de la performance réelle.
- `chmax` a un effet positif et significatif (p-value = 0.037) :
    - Une augmentation du nombre maximal de canaux est associée à une augmentation de la performance.
- `cach` a un effet positif mais seulement faiblement significatif (p-value = 0.080).
- Les variables `syct`, `mmin`, `mmax`, `chmin` ne sont pas significatives individuellement.

Nous affinerons enfin le modèle en sélectionnant automatiquement les variables pertinentes avec la fonction `step` et la direction `backward` :

```
Start:  AIC=1555.64
perf ~ syct + mmin + mmax + cach + chmin + chmax + estperf

          Df Sum of Sq    RSS    AIC
- chmin    1        57 330773 1553.7
- syct     1      1903 332619 1554.8
- mmax     1      2020 332736 1554.9
- mmin     1      2275 332991 1555.1
<none>                 330716 1555.6
- cach     1      5081 335797 1556.8
- chmax    1      7246 337963 1558.2
- estperf  1    396286 727002 1718.3
```

```
Step:  AIC=1553.67
perf ~ syct + mmin + mmax + cach + chmax + estperf

          Df Sum of Sq    RSS    AIC
- syct     1      1881 332654 1552.9
- mmax     1      2057 332830 1553.0
- mmin     1      2219 332992 1553.1
<none>                 330773 1553.7
- cach     1      5147 335920 1554.9
- chmax    1      7545 338318 1556.4
- estperf  1    396587 727360 1716.4

Step:  AIC=1552.86
perf ~ mmin + mmax + cach + chmax + estperf

          Df Sum of Sq    RSS    AIC
- mmax     1      1131 333785 1551.6
<none>                 332654 1552.9
- mmin     1      3456 336110 1553.0
- cach     1      6747 339400 1555.0
- chmax    1      9281 341935 1556.6
- estperf  1    423060 755713 1722.3

Step:  AIC=1551.57
perf ~ mmin + cach + chmax + estperf

          Df Sum of Sq     RSS    AIC
- mmin     1      3115  336900 1551.5
<none>                 333785 1551.6
- cach     1      7771  341555 1554.4
- chmax    1      8975  342760 1555.1
- estperf  1    674856 1008640 1780.7

Step:  AIC=1551.51
perf ~ cach + chmax + estperf

          Df Sum of Sq     RSS    AIC
<none>                 336900 1551.5
- chmax    1      5998  342897 1553.2
- cach     1      8884  345783 1555.0
- estperf  1   2119750 2456650 1964.7

Call:
lm(formula = perf ~ cach + chmax + estperf, data = cpus_quant)

Residuals:
     Min       1Q   Median       3Q      Max
-162.986  -14.246   -3.952    8.759  235.691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.89097    3.56528   0.530   0.5964
```

```
cach         0.21443    0.09223   2.325   0.0210 *
chmax        0.26008    0.13614   1.910   0.0575 .
estperf      0.94201    0.02623  35.914   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.54 on 205 degrees of freedom
Multiple R-squared:  0.9374,    Adjusted R-squared:  0.9365
F-statistic:  1023 on 3 and 205 DF,  p-value: < 2.2e-16
```

Commentaire

La procédure de sélection pas à pas basée sur le critère AIC conduit à un modèle parcimonieux ne retenant que **3 variables explicatives** : `cach`, `chmax` et `estperf` :

- Le modèle est **globalement très significatif** (F-statistic = 1023, p-value < 2.2e-16).
- `estperf` a un coefficient très élevé et extrêmement significatif (p < 2e-16).
  - C'est de loin la variable la plus explicative du modèle.
  - Une augmentation de 1 unité de `estperf` entraîne en moyenne une augmentation d'environ **0.94 unités de** `perf`, toutes choses égales par ailleurs.
- `cach`
  - Son coefficient indique que les processeurs disposant d'une plus grande mémoire cache ont de meilleures performances, toutes choses égales par ailleurs.
- `chmax`
  - Son coefficient indique que le nombre maximal de canaux contribue positivement à la performance, mais de manière plus modérée.

En comparaison avec le modèle complet, le modèle sélectionné explique **presque autant de variance** et ce, avec **moins de variables**.

Exercice 22.1 Données `enseignants`

Ces données sont extraites d'un reccueil de données issu d'une enquête portant sur une population d'enseignants de collèges.

Elles ont été modifiées pour les besoins du TP. La plupart des variables sont explicites.

Le salaire est exprimé en `euros`, l'âge et l'ancienneté en `années`. Le stress, l'estime de soi et la satisfaction au travail sont mesurés sur des échelles allant de `0 à 50` suivant des techniques appropriées.

Importons les données :

Table 2.16 – Aperçu des données `enseignants`

Sexe	Age	EtatCivil	Nbenfant	Diplome	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	AvisReforme
Homme	37	Célibataire	0	Bac+3	11	1600	14.45	15.70	16.15	Défavorable
Homme	38	Célibataire	2	Bac+3	14	1670	17.57	18.88	17.56	Défavorable
Femme	29	Célibataire	0	Bac+3	1	1600	4.05	21.38	4.31	Défavorable
Homme	53	Marié(e)	2	Bac+3	28	1896	32.55	13.88	34.56	Défavorable
Homme	30	Marié(e)	1	Bac+3	7	1996	10.50	17.90	10.05	Défavorable
Homme	44	Marié(e)	2	Bac+3	18	1960	22.16	18.76	22.62	Défavorable

Exercice 22.2. Résumé statistique des données

Donnons un résumé statistique descriptif standard des données avec les fonctions `str` et `summary` :

Table 2.17 – Structure des données

Variable	Type	Valeurs
Sexe	chr	: chr [1 :168] ”Homme” ”Homme” ”Femme” ”Homme” ...
Age	num	: num [1 :168] 37 38 29 53 30 44 43 25 37 40 ...
EtatCivil	chr	: chr [1 :168] ”Célibataire” ”Célibataire” ”Célibataire” ”Marié(e)” ...
Nbenfant	num	: num [1 :168] 0 2 0 2 1 2 1 0 0 2 ...
Diplome	chr	: chr [1 :168] ”Bac+3” ”Bac+3” ”Bac+3” ”Bac+3” ...
Anciennete	num	: num [1 :168] 11 14 1 28 7 18 16 1 13 14 ...
Salaire	num	: num [1 :168] 1600 1670 1600 1896 1996 ...
Satisfaction	num	$ Satisfaction : num [1 :168] 14.45 17.57 4.05 32.55 10.5 ...
Stress	num	: num [1 :168] 15.7 18.9 21.4 13.9 17.9 ...
EstimeSoi	num	: num [1 :168] 16.15 17.56 4.31 34.56 10.05 ...
AvisReforme	chr	$ AvisReforme : chr [1 :168] ”Défavorable” ”Défavorable” ”Défavorable” ”Défavorable” ...

Table 2.18 – Tri à plat de la variable Sexe

Modalités	Effectifs
Femme	53
Homme	115

Table 2.19 – Tri à plat de la variable EtatCivil

Modalités	Effectifs
Célibataire	24
Divorcé(e)	13
Marié(e)	124
Veuf(ve)	7

Table 2.20 – Tri à plat de la variable Diplome

Modalités	Effectifs
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6
Bac+7	1

Table 2.21 – Tri à plat de la variable AvisReforme

Modalités	Effectifs
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

Table 2.22 – Tri à plat de la variable Nbenfant

Modalités	Effectifs
0	33
1	22
2	77
3	32
4	3
5	1

Table 2.23 – Résumé statistique de Age

Statistique	Valeur
Min.	25.00
1st Qu.	37.00
Median	41.00
Mean	41.99
3rd Qu.	49.25
Max.	57.00

Table 2.24 – Résumé statistique de Anciennete

Statistique	Valeur
Min.	1.00
1st Qu.	10.00
Median	15.00
Mean	16.55
3rd Qu.	24.25
Max.	34.00

Table 2.25 – Résumé statistique de Salaire

Statistique	Valeur
Min.	1200.0

1st Qu.	1650.0
Median	1720.0
Mean	1777.9
3rd Qu.	1907.5
Max.	2200.0

Table 2.26 – Résumé statistique de Satisfaction

Statistique	Valeur
Min.	3.85
1st Qu.	13.84
Median	19.17
Mean	20.43
3rd Qu.	28.31
Max.	38.45

Table 2.27 – Résumé statistique de Stress

Statistique	Valeur
Min.	3.70
1st Qu.	15.19
Median	18.19
Mean	18.20
3rd Qu.	21.11
Max.	31.84

Table 2.28 – Résumé statistique de EstimeSoi

Statistique	Valeur
Min.	3.54
1st Qu.	14.03
Median	19.69
Mean	21.08
3rd Qu.	29.83
Max.	42.15

Exercice 22.3. Analyse des résumés

Examinons les résumés et en particulier celui du salaire :

- Le salaire des enseignants interrogés s'étend de **1200 € à 2200 €**, ce qui indique une dispersion modérée des rémunérations.
    - La **médiane** est de **1720 €**, valeur proche de la **moyenne (1778 €)**, ce qui suggère une distribution globalement assez symétrique, sans forte asymétrie marquée.

    - Le **premier quartile (1650 €)** et le **troisième quartile (1908 €)** montrent que 50 % des salaires se situent dans un intervalle relativement resserré d'environ 260 €, traduisant une certaine homogénéité salariale au sein de l'échantillon.
    - Les valeurs extrêmes restent limitées, ce qui laisse supposer l'absence de salaires aberrants ou très atypiques.
- Cette distribution est cohérente avec les autres variables liées à la carrière professionnelle :
    - une **ancienneté moyenne de 16,55 ans**,
    - une majorité de diplômes situés entre **Bac+3 et Bac+4**,
    - et un effectif principalement composé d'enseignants mariés, donc probablement en milieu ou fin de carrière.

En conclusion, le salaire apparaît comme une variable relativement stable dans cet échantillon, avec des variations plausibles et cohérentes avec l'âge, l'ancienneté et le niveau de diplôme.

Exercice 22.4. Croisement qualitatif vs qualitatif

- On donnera les tableaux de contingences (effectifs, fréquences, pourcentages)
- On donnera aussi différentes représentations graphiques du tableau de contingence des effectifs : `balloonplot`, `barplot` avec les options `beside=TRUE` et `beside=FALSE`, `mosaicplot`
- On donnera les distributions marginales et on remarquera qu'elles sont, en fait, pareilles aux distributions univariées.
- On y ajoutera les distributions conditionnelles.

Table 2.29 – Tableau de contingence (effectifs) : Sexe vs EtatCivil

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
Femme	7	3	38	5
Homme	17	10	86	2

Table 2.30 – Tableau de contingence (fréquences) : Sexe vs EtatCivil

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
Femme	0.042	0.018	0.226	0.030
Homme	0.101	0.060	0.512	0.012

Table 2.31 – Tableau de contingence (pourcentages) : Sexe vs EtatCivil

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
Femme	4.2	1.8	22.6	3.0
Homme	10.1	6.0	51.2	1.2

Table 2.32 – Distribution marginale : Sexe

Var1	Freq
Femme	53
Homme	115

Table 2.33 – Distribution marginale : EtatCivil

Var1	Freq
Célibataire	24
Divorcé(e)	13
Marié(e)	124
Veuf(ve)	7

Table 2.34 – Distribution marginale : Sexe

Sexe	Freq
Femme	53
Homme	115

Table 2.35 – Distribution marginale : EtatCivil

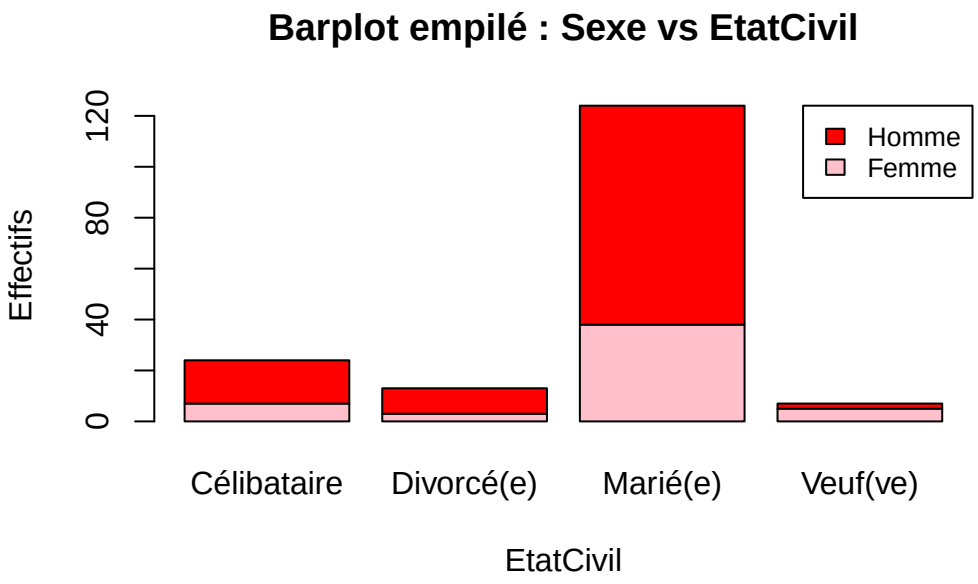
EtatCivil	Freq
Célibataire	24
Divorcé(e)	13
Marié(e)	124
Veuf(ve)	7

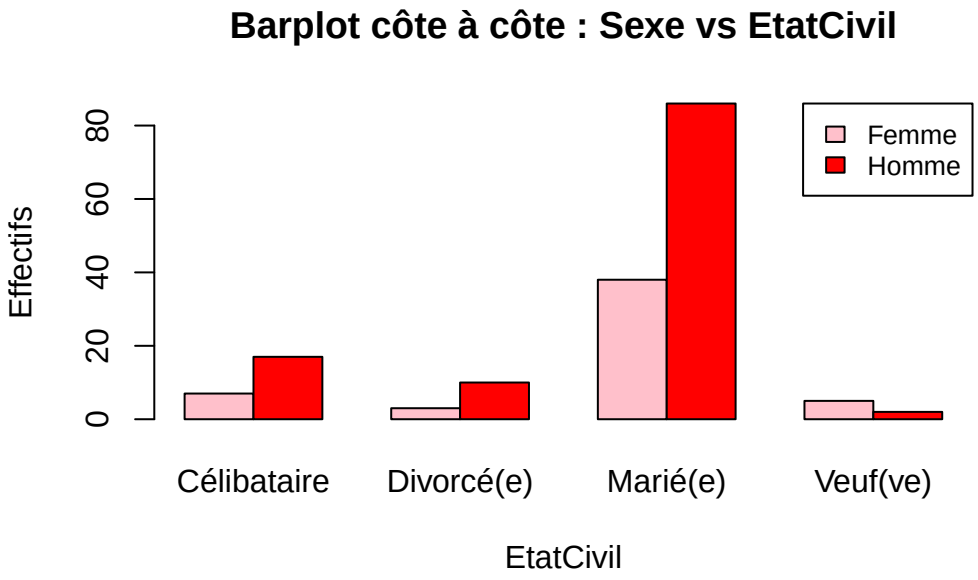
Table 2.36 – Distribution conditionnelle : Sexe sachant EtatCivil

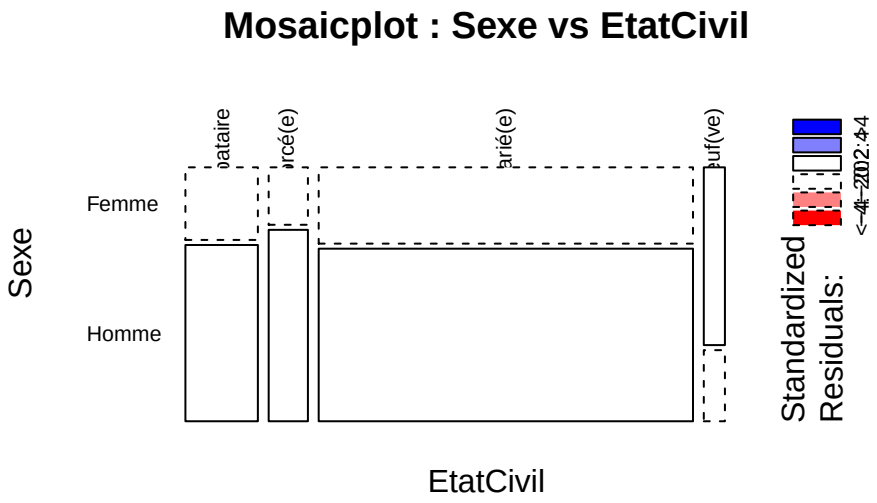
	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
Femme	0.2916667	0.2307692	0.3064516	0.7142857
Homme	0.7083333	0.7692308	0.6935484	0.2857143

Table 2.37 – Distribution conditionnelle : EtatCivil sachant Sexe

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
Femme	0.1320755	0.0566038	0.7169811	0.0943396
Homme	0.1478261	0.0869565	0.7478261	0.0173913


Barplot empilé : Sexe vs EtatCivil
Effectifs
0
40
80
120
Homme
Femme
Célibataire
Divorcé(e)
Marié(e)
Veuf(ve)
EtatCivil


Barplot côte à côte : Sexe vs EtatCivil
Effectifs
0
20
40
60
80
Femme
Homme
Célibataire
Divorcé(e)
Marié(e)
Veuf(ve)
EtatCivil



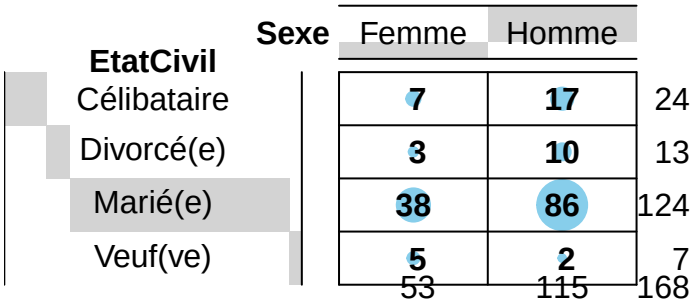


Table 2.38 – Tableau de contingence (effectifs) : Sexe vs Diplome

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Femme	0	9	39	1	3	1
Homme	1	33	78	0	3	0

Table 2.39 – Tableau de contingence (fréquences) : Sexe vs Diplome

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Femme	0.000	0.054	0.232	0.006	0.018	0.006
Homme	0.006	0.196	0.464	0.000	0.018	0.000

Table 2.40 – Tableau de contingence (pourcentages) : Sexe vs Diplome

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Femme	0.0	5.4	23.2	0.6	1.8	0.6
Homme	0.6	19.6	46.4	0.0	1.8	0.0

Table 2.41 – Distribution marginale : Sexe

Var1	Freq
Femme	53
Homme	115

Table 2.42 – Distribution marginale : Diplome

Var1	Freq
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6
Bac+7	1

Table 2.43 – Distribution marginale : Sexe

Sexe	Freq
Femme	53
Homme	115

Table 2.44 – Distribution marginale : Diplome

Diplome	Freq
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6

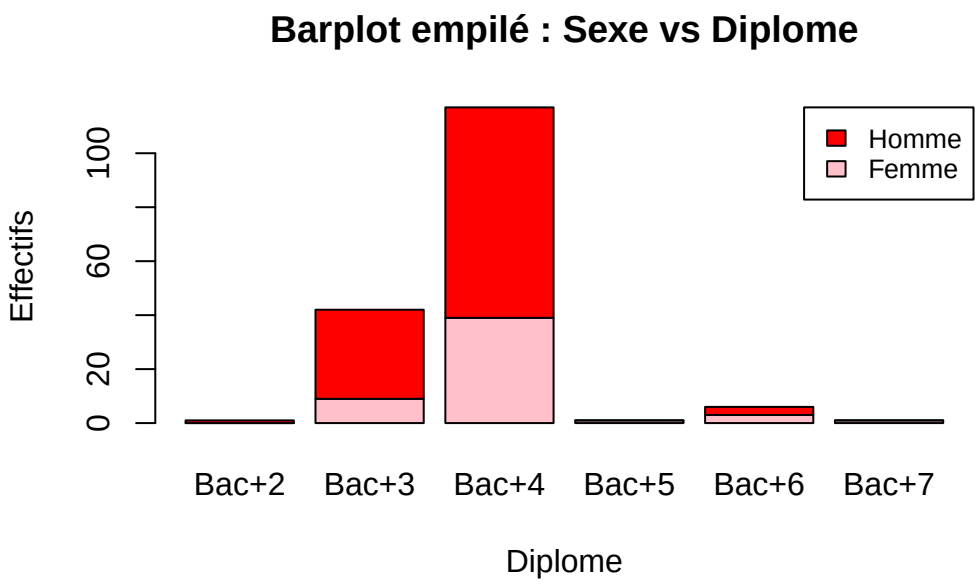
Diplome	Freq
Bac+7	1

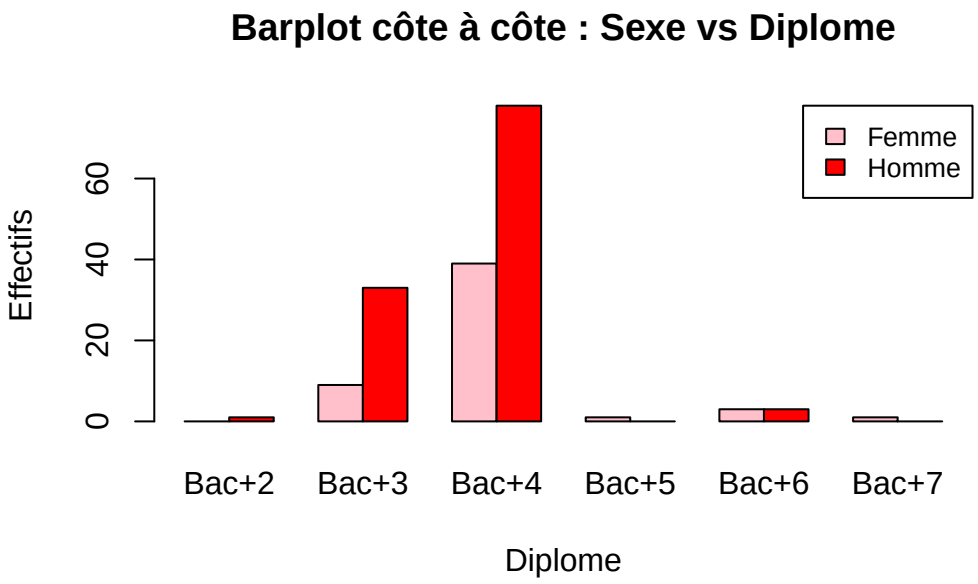
Table 2.45 – Distribution conditionnelle : Sexe sachant Diplome

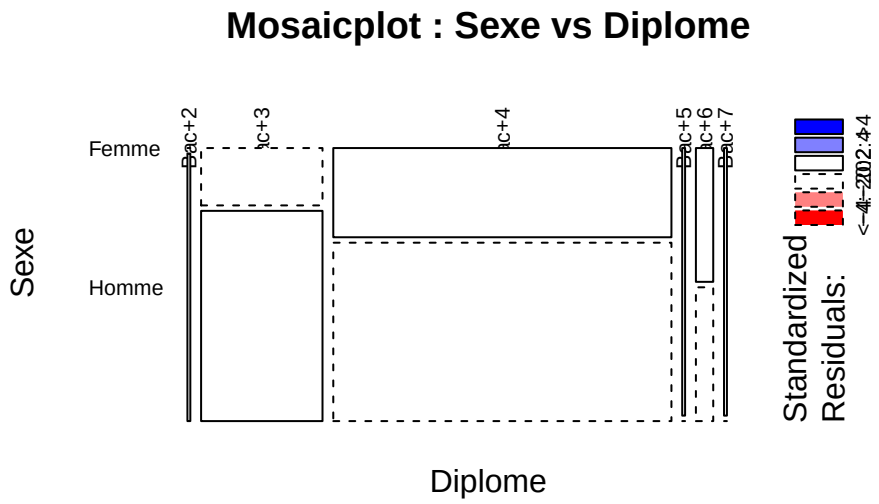
	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Femme	0	0.2142857	0.3333333	1	0.5	1
Homme	1	0.7857143	0.6666667	0	0.5	0

Table 2.46 – Distribution conditionnelle : Diplome sachant Sexe

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Femme	0.0000000	0.1698113	0.7358491	0.0188679	0.0566038	0.0188679
Homme	0.0086957	0.2869565	0.6782609	0.0000000	0.0260870	0.0000000




Barplot côte à côte : Sexe vs Diplome
Femme
Homme
Effectifs
0
20
40
60
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7
Diplome


Mosaicplot : Sexe vs Diplome
Femme
Homme
Sexe
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7
Standardized Residuals:
Diplome

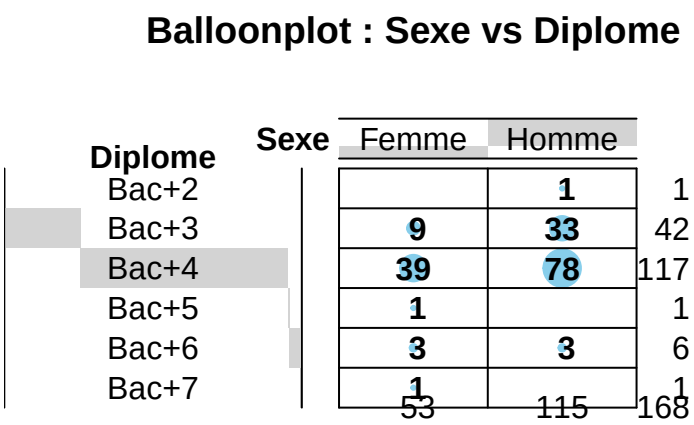


Table 2.47 – Tableau de contingence (effectifs) : Sexe vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Femme	23	1	9	4	16
Homme	62	7	18	9	19

Table 2.48 – Tableau de contingence (fréquences) : Sexe vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Femme	0.137	0.006	0.054	0.024	0.095
Homme	0.369	0.042	0.107	0.054	0.113

Table 2.49 – Tableau de contingence (pourcentages) : Sexe vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Femme	13.7	0.6	5.4	2.4	9.5
Homme	36.9	4.2	10.7	5.4	11.3

Table 2.50 – Distribution marginale : Sexe

Var1	Freq
Femme	53

Var1	Freq
Homme	115

Table 2.51 – Distribution marginale : AvisReforme

Var1	Freq
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

Table 2.52 – Distribution marginale : Sexe

Sexe	Freq
Femme	53
Homme	115

Table 2.53 – Distribution marginale : AvisReforme

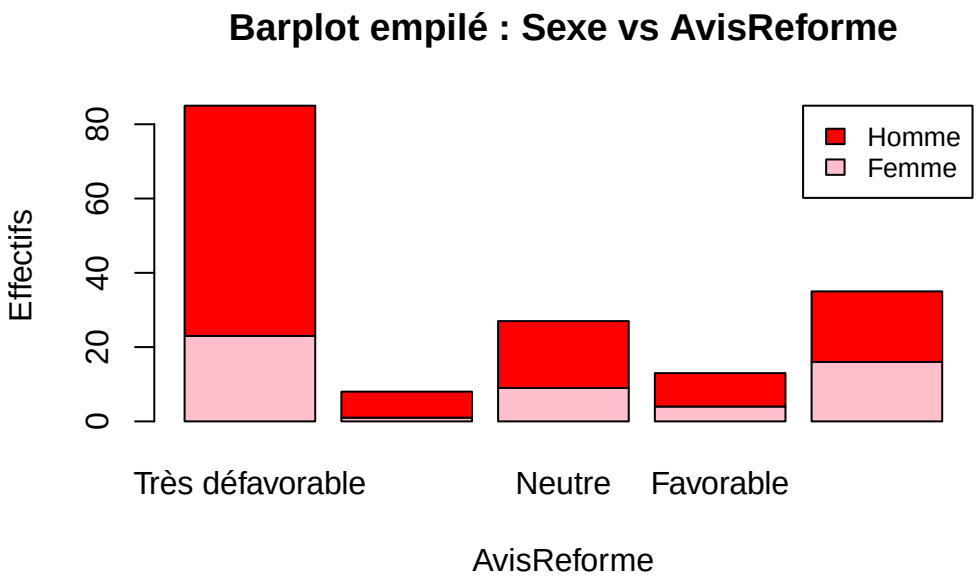
AvisReforme	Freq
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

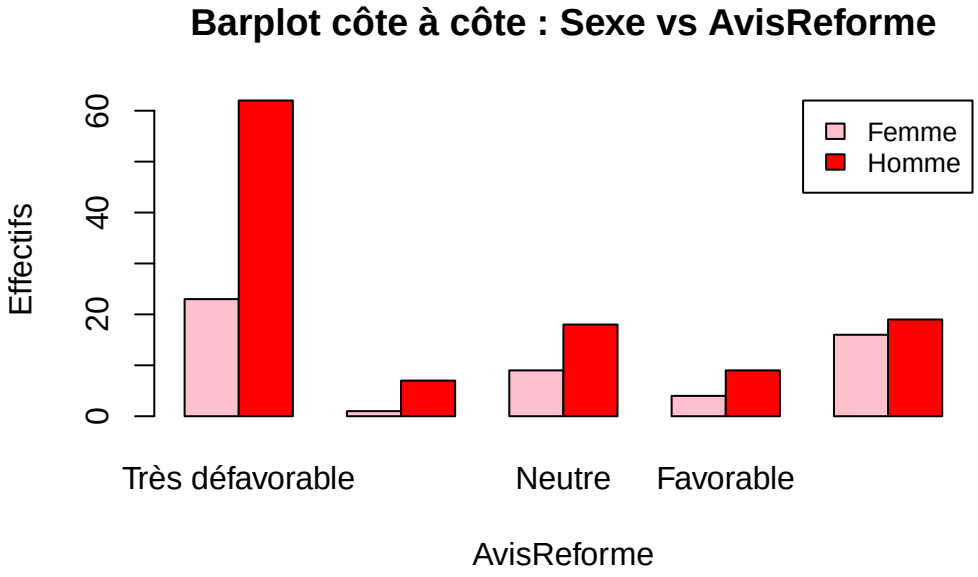
Table 2.54 – Distribution conditionnelle : Sexe sachant AvisReforme

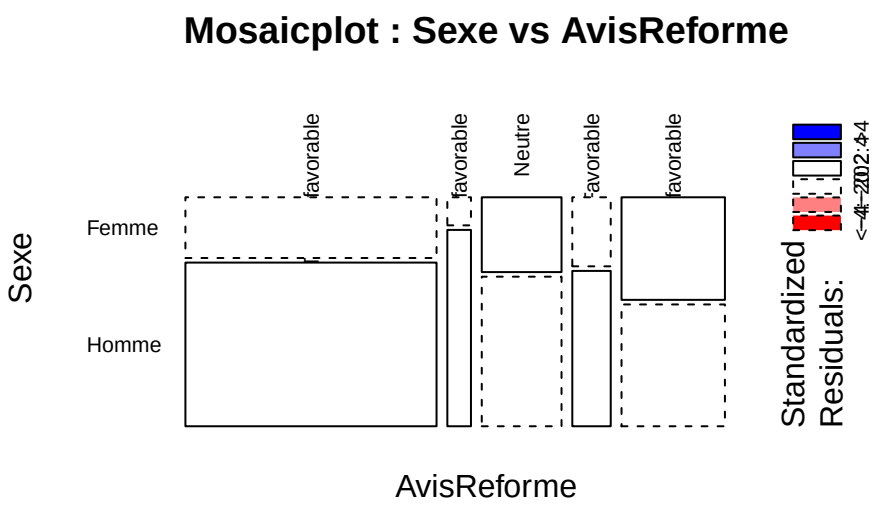
	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Femme	0.2705882	0.125	0.3333333	0.3076923	0.4571429
Homme	0.7294118	0.875	0.6666667	0.6923077	0.5428571

Table 2.55 – Distribution conditionnelle : AvisReforme sachant Sexe

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Femme	0.4339623	0.0188679	0.1698113	0.0754717	0.3018868
Homme	0.5391304	0.0608696	0.1565217	0.0782609	0.1652174


Barplot empilé : Sexe vs AvisReforme
Effectifs
0
20
40
60
80
Homme
Femme
Très défavorable
Neutre
Favorable
AvisReforme


Barplot côte à côte : Sexe vs AvisReforme
Effectifs
0
20
40
60
Femme
Homme
Très défavorable
Neutre
Favorable
AvisReforme



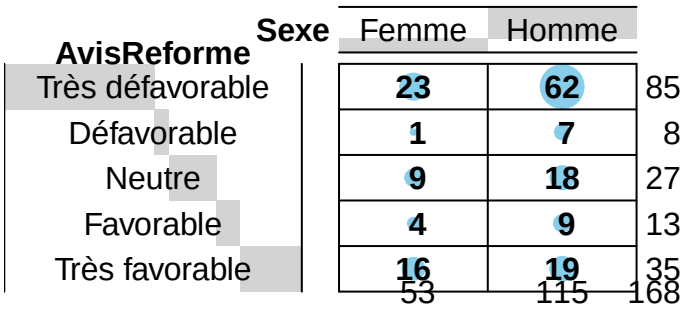


Table 2.56 – Tableau de contingence (effectifs) : EtatCivil vs Diplome

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Célibataire	0	22	0	1	1	0
Divorcé(e)	0	11	0	0	2	0
Marié(e)	1	9	110	0	3	1

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Veuf(ve)	0	0	7	0	0	0

Table 2.57 – Tableau de contingence (fréquences) : EtatCivil vs Diplome

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Célibataire	0.000	0.131	0.000	0.006	0.006	0.000
Divorcé(e)	0.000	0.065	0.000	0.000	0.012	0.000
Marié(e)	0.006	0.054	0.655	0.000	0.018	0.006
Veuf(ve)	0.000	0.000	0.042	0.000	0.000	0.000

Table 2.58 – Tableau de contingence (pourcentages) : EtatCivil vs Diplome

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Célibataire	0.0	13.1	0.0	0.6	0.6	0.0
Divorcé(e)	0.0	6.5	0.0	0.0	1.2	0.0
Marié(e)	0.6	5.4	65.5	0.0	1.8	0.6
Veuf(ve)	0.0	0.0	4.2	0.0	0.0	0.0

Table 2.59 – Distribution marginale : EtatCivil

Var1	Freq
Célibataire	24
Divorcé(e)	13
Marié(e)	124
Veuf(ve)	7

Table 2.60 – Distribution marginale : Diplome

Var1	Freq
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6
Bac+7	1

Table 2.61 – Distribution marginale : EtatCivil

EtatCivil	Freq
Célibataire	24

EtatCivil	Freq
Divorcé(e)	13
Marié(e)	124
Veuf(ve)	7

Table 2.62 – Distribution marginale : Diplome

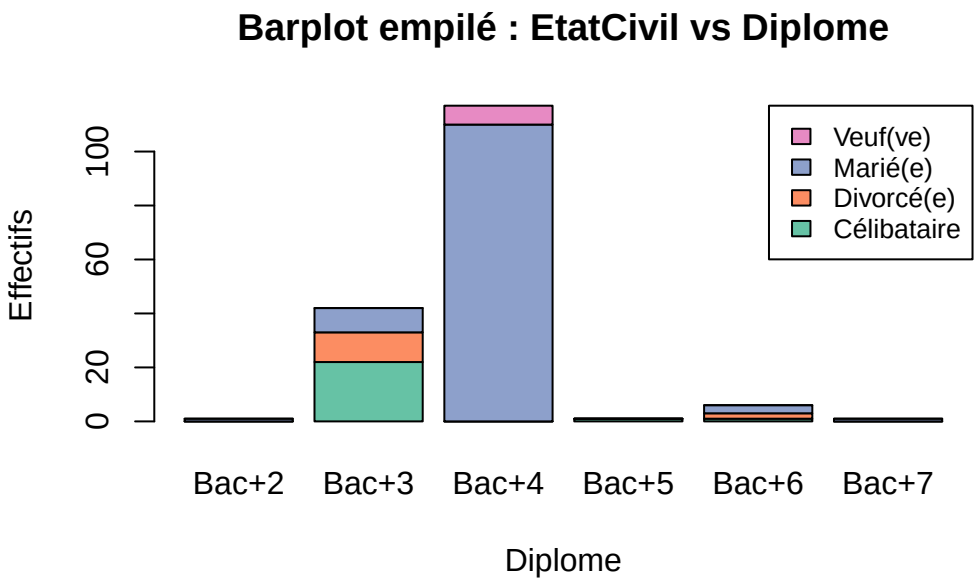
Diplome	Freq
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6
Bac+7	1

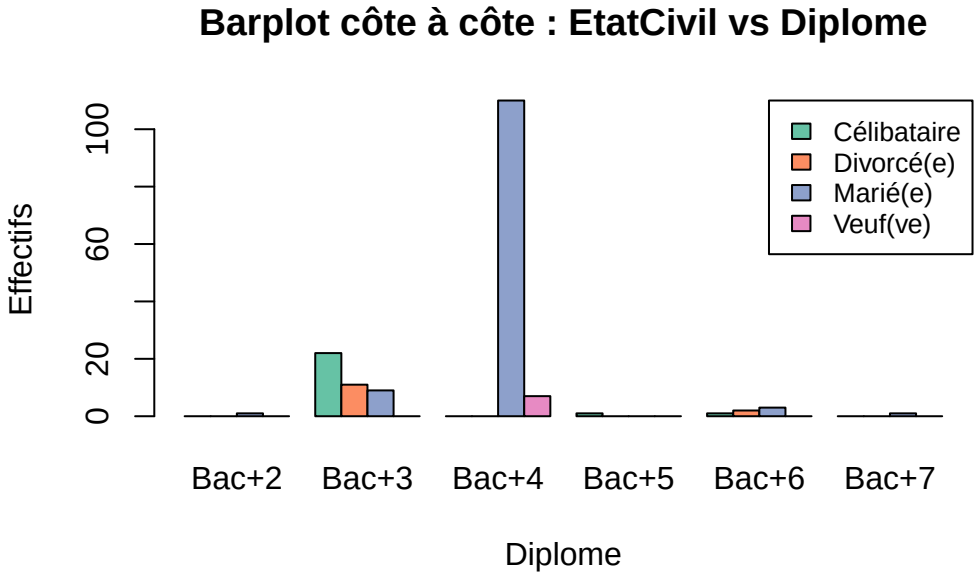
Table 2.63 – Distribution conditionnelle : EtatCivil sachant Diplome

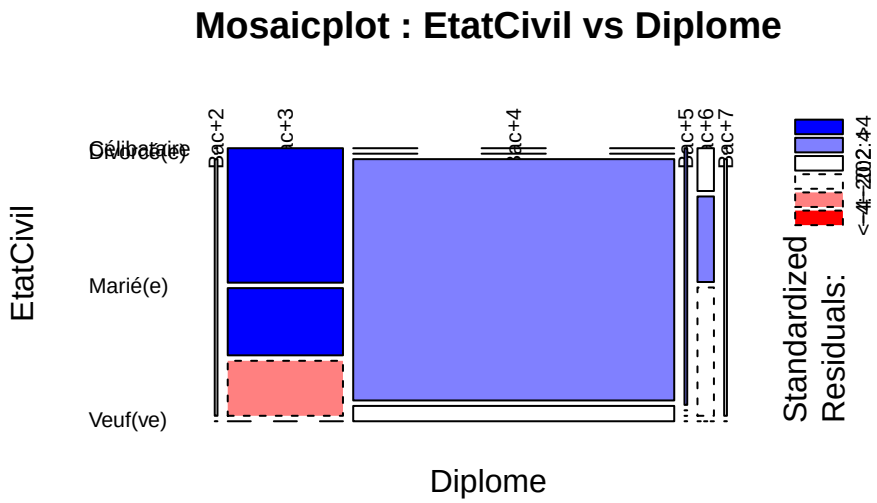
	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Célibataire	0	0.5238095	0.0000000	1	0.1666667	0
Divorcé(e)	0	0.2619048	0.0000000	0	0.3333333	0
Marié(e)	1	0.2142857	0.9401709	0	0.5000000	1
Veuf(ve)	0	0.0000000	0.0598291	0	0.0000000	0

Table 2.64 – Distribution conditionnelle : Diplome sachant EtatCivil

	Bac+2	Bac+3	Bac+4	Bac+5	Bac+6	Bac+7
Célibataire	0.0000000	0.9166667	0.0000000	0.0416667	0.0416667	0.0000000
Divorcé(e)	0.0000000	0.8461538	0.0000000	0.0000000	0.1538462	0.0000000
Marié(e)	0.0080645	0.0725806	0.8870968	0.0000000	0.0241935	0.0080645
Veuf(ve)	0.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000


Barplot empilé : EtatCivil vs Diplome
Effectifs
0
20
60
100
Veuf(ve)
Marié(e)
Divorcé(e)
Célibataire
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7
Diplome


Barplot côte à côte : EtatCivil vs Diplome
Effectifs
0
20
60
100
Célibataire
Divorcé(e)
Marié(e)
Veuf(ve)
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7
Diplome



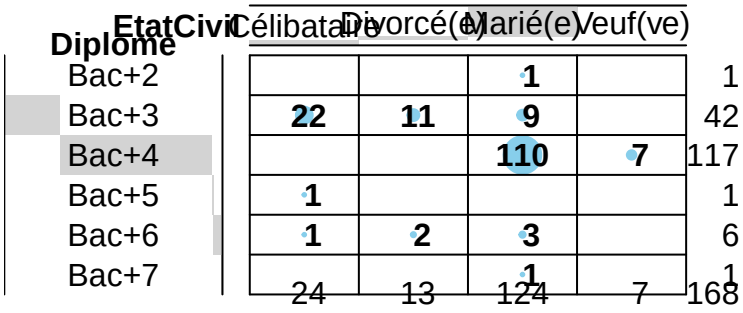


Table 2.65 – Tableau de contingence (effectifs) : EtatCivil vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Célibataire	17	4	0	0	3
Divorcé(e)	9	0	2	0	2
Marié(e)	56	4	25	12	27

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Veuf(ve)	3	0	0	1	3

Table 2.66 – Tableau de contingence (fréquences) : EtatCivil vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Célibataire	0.101	0.024	0.000	0.000	0.018
Divorcé(e)	0.054	0.000	0.012	0.000	0.012
Marié(e)	0.333	0.024	0.149	0.071	0.161
Veuf(ve)	0.018	0.000	0.000	0.006	0.018

Table 2.67 – Tableau de contingence (pourcentages) : EtatCivil vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Célibataire	10.1	2.4	0.0	0.0	1.8
Divorcé(e)	5.4	0.0	1.2	0.0	1.2
Marié(e)	33.3	2.4	14.9	7.1	16.1
Veuf(ve)	1.8	0.0	0.0	0.6	1.8

Table 2.68 – Distribution marginale : EtatCivil

Var1	Freq
Célibataire	24
Divorcé(e)	13
Marié(e)	124
Veuf(ve)	7

Table 2.69 – Distribution marginale : AvisReforme

Var1	Freq
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

Table 2.70 – Distribution marginale : EtatCivil

EtatCivil	Freq
Célibataire	24
Divorcé(e)	13

EtatCivil	Freq
Marié(e)	124
Veuf(ve)	7

Table 2.71 – Distribution marginale : AvisReforme

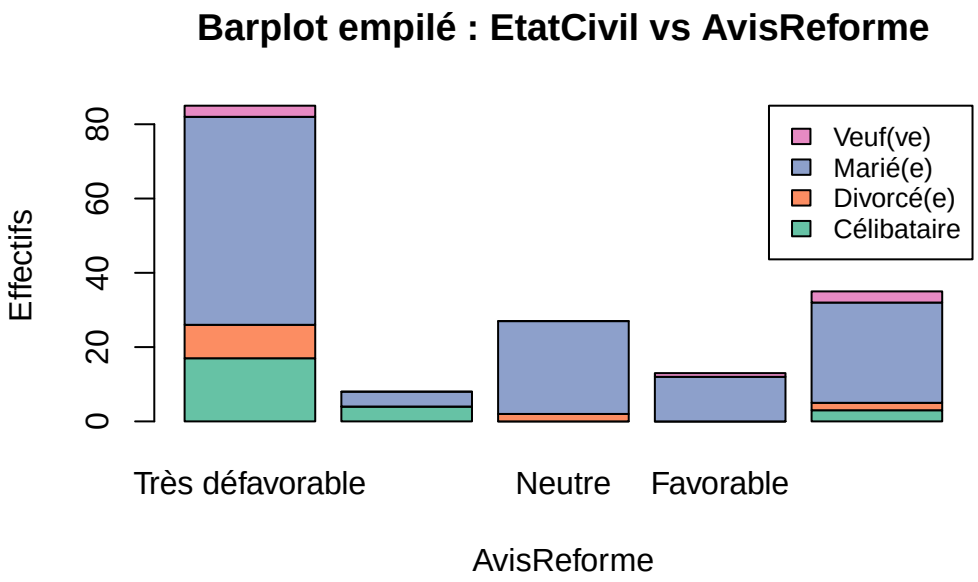
AvisReforme	Freq
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

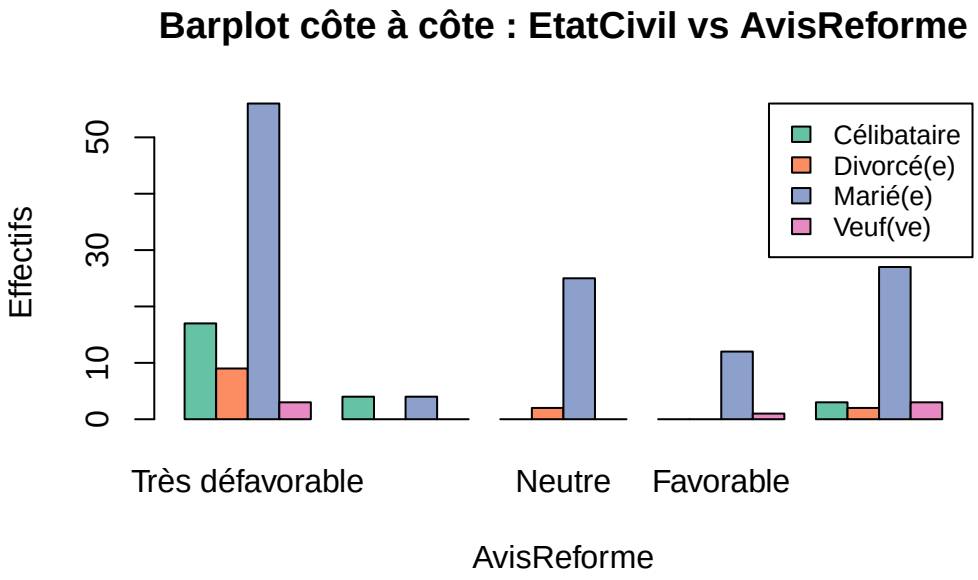
Table 2.72 – Distribution conditionnelle : EtatCivil sachant AvisReforme

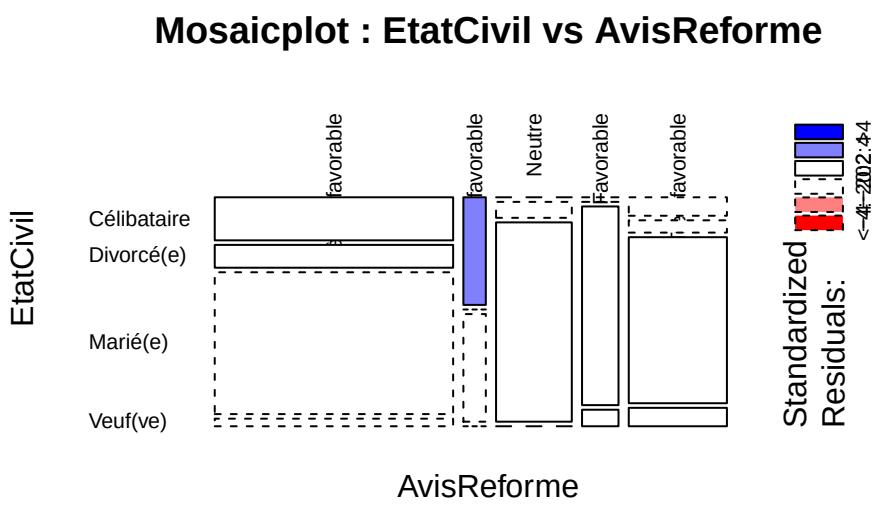
	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Célibataire	0.2000000	0.5	0.0000000	0.0000000	0.0857143
Divorcé(e)	0.1058824	0.0	0.0740741	0.0000000	0.0571429
Marié(e)	0.6588235	0.5	0.9259259	0.9230769	0.7714286
Veuf(ve)	0.0352941	0.0	0.0000000	0.0769231	0.0857143

Table 2.73 – Distribution conditionnelle : AvisReforme sachant EtatCivil

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Célibataire	0.7083333	0.1666667	0.0000000	0.0000000	0.1250000
Divorcé(e)	0.6923077	0.0000000	0.1538462	0.0000000	0.1538462
Marié(e)	0.4516129	0.0322581	0.2016129	0.0967742	0.2177419
Veuf(ve)	0.4285714	0.0000000	0.0000000	0.1428571	0.4285714


Barplot empilé : EtatCivil vs AvisReforme
Effectifs
0
20
40
60
80
Veuf(ve)
Marié(e)
Divorcé(e)
Célibataire
Très défavorable
Neutre
Favorable
AvisReforme


Barplot côte à côte : EtatCivil vs AvisReforme
Effectifs
0
10
30
50
Célibataire
Divorcé(e)
Marié(e)
Veuf(ve)
Très défavorable
Neutre
Favorable
AvisReforme



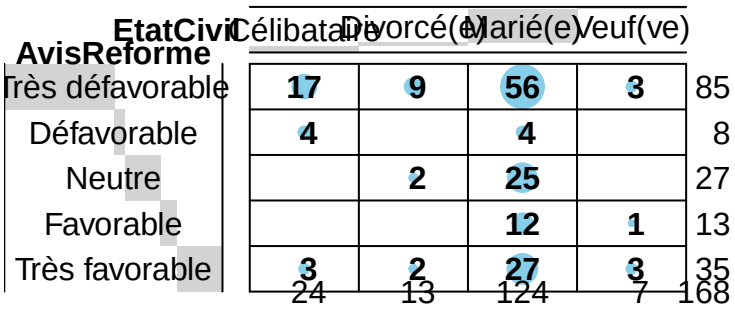


Table 2.74 – Tableau de contingence (effectifs) : Diplome vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Bac+2	0	0	0	0	1
Bac+3	24	7	2	4	5
Bac+4	57	0	24	9	27

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Bac+5	1	0	0	0	0
Bac+6	3	1	1	0	1
Bac+7	0	0	0	0	1

Table 2.75 – Tableau de contingence (fréquences) : Diplome vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Bac+2	0.000	0.000	0.000	0.000	0.006
Bac+3	0.143	0.042	0.012	0.024	0.030
Bac+4	0.339	0.000	0.143	0.054	0.161
Bac+5	0.006	0.000	0.000	0.000	0.000
Bac+6	0.018	0.006	0.006	0.000	0.006
Bac+7	0.000	0.000	0.000	0.000	0.006

Table 2.76 – Tableau de contingence (pourcentages) : Diplome vs AvisReforme

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Bac+2	0.0	0.0	0.0	0.0	0.6
Bac+3	14.3	4.2	1.2	2.4	3.0
Bac+4	33.9	0.0	14.3	5.4	16.1
Bac+5	0.6	0.0	0.0	0.0	0.0
Bac+6	1.8	0.6	0.6	0.0	0.6
Bac+7	0.0	0.0	0.0	0.0	0.6

Table 2.77 – Distribution marginale : Diplome

Var1	Freq
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6
Bac+7	1

Table 2.78 – Distribution marginale : AvisReforme

Var1	Freq
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

Table 2.79 – Distribution marginale : Diplome

Diplome	Freq
Bac+2	1
Bac+3	42
Bac+4	117
Bac+5	1
Bac+6	6
Bac+7	1

Table 2.80 – Distribution marginale : AvisReforme

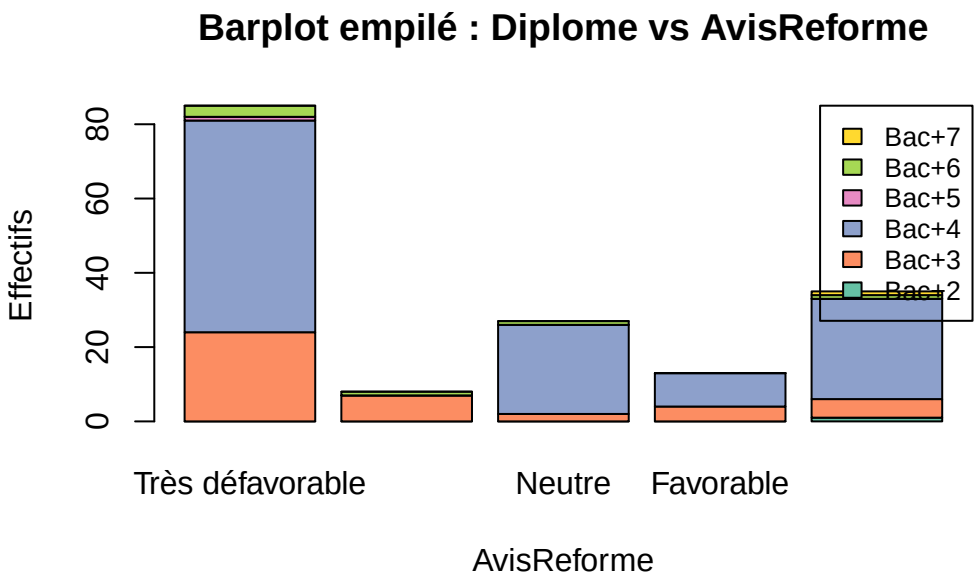
AvisReforme	Freq
Très défavorable	85
Défavorable	8
Neutre	27
Favorable	13
Très favorable	35

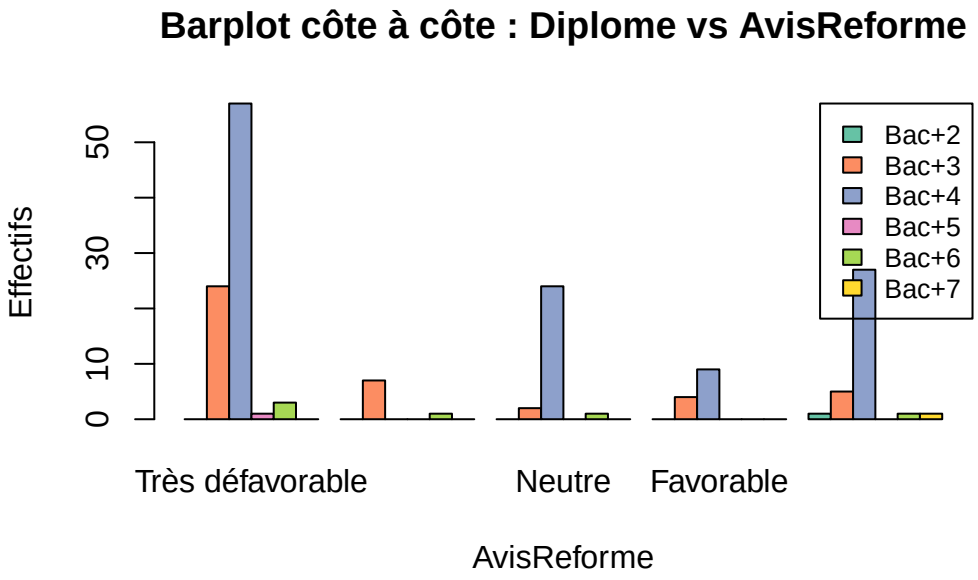
Table 2.81 – Distribution conditionnelle : Diplome sachant AvisReforme

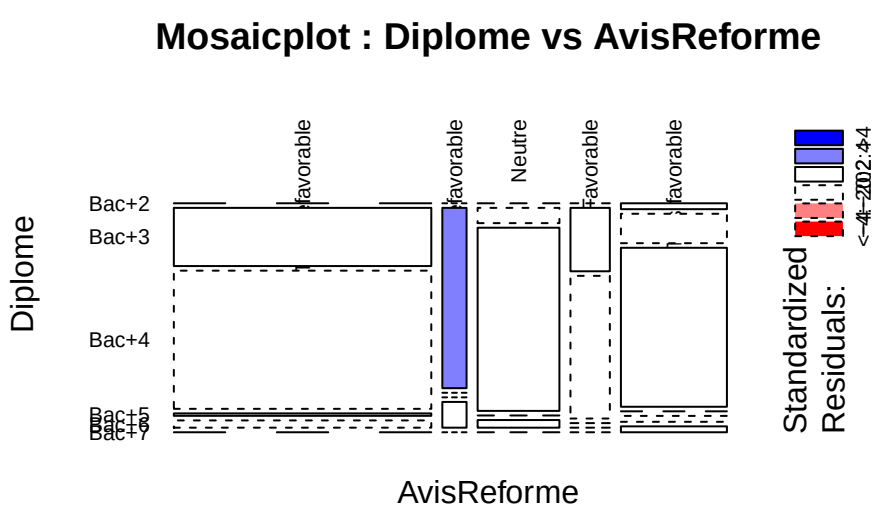
	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Bac+2	0.0000000	0.000	0.0000000	0.0000000	0.0285714
Bac+3	0.2823529	0.875	0.0740741	0.3076923	0.1428571
Bac+4	0.6705882	0.000	0.8888889	0.6923077	0.7714286
Bac+5	0.0117647	0.000	0.0000000	0.0000000	0.0000000
Bac+6	0.0352941	0.125	0.0370370	0.0000000	0.0285714
Bac+7	0.0000000	0.000	0.0000000	0.0000000	0.0285714

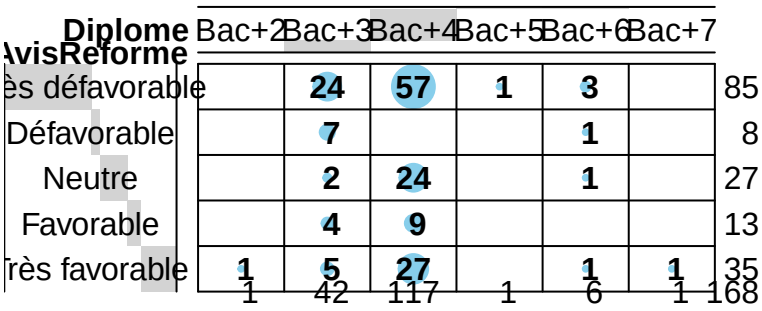
Table 2.82 – Distribution conditionnelle : AvisReforme sachant Diplome

	Très défavorable	Défavorable	Neutre	Favorable	Très favorable
Bac+2	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000
Bac+3	0.5714286	0.1666667	0.0476190	0.0952381	0.1190476
Bac+4	0.4871795	0.0000000	0.2051282	0.0769231	0.2307692
Bac+5	1.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Bac+6	0.5000000	0.1666667	0.1666667	0.0000000	0.1666667
Bac+7	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000


Barplot empilé : Diplome vs AvisReforme
Effectifs
0
20
40
60
80
Bac+7
Bac+6
Bac+5
Bac+4
Bac+3
Bac+2
Très défavorable
Neutre
Favorable
AvisReforme


Barplot côte à côte : Diplome vs AvisReforme
Effectifs
0
10
30
50
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7
Très défavorable
Neutre
Favorable
AvisReforme





e) Faisons maintenant un test de χ^2 (avec la fonction `chisq.test`) afin d'apprécier la dépendance, ou non, des variables `Sexe` et `EtatCivil` :

```
    Pearson's Chi-squared test

data:  tab_obs
X-squared = 5.6972, df = 3, p-value = 0.1273
```

- La p-valeur résultant du test de χ^2 est bien supérieure à 0.05. Par conséquent, on **conserve** l'hypothèse nulle d'indépendance.

f) Récupérons le tableau correspondant à l'indépendance de la question précédente.

Calculons le χ^2 par étapes (Tableau des effectifs théoriques correspondant à l'indépendance des variables, Calcul du coefficient de chi-deux par étapes, ...) et comparons aux résultats obtenus à la question précédente :

```
Chi2 observée :  5.697248

Chi2 manuel:  5.697248
```

g) Testons et concluons de deux manières (à partir du coefficient et à partir de la p-valeur) sur la liaision entre les variables `Sexe` et `EtatCivil` :

```
Chi2 observée :  5.697248

Chi² critique (5%)  : 7.814728

p-value             : 0.1273056
```

- **Décision basée sur chi2**
  - La statistique de χ^2 observée est inférieure à la χ^{2*} critique. Par conséquent, on **conserve** l'hypothèse nulle d'indépendance.
- **Décision basée sur la p-valeur**
  - La p-valeur résultant du test de χ^2 est bien supérieure à 0.05. Par conséquent, on **conserve** l'hypothèse nulle d'indépendance.

Exercice 22.5. Croisement quantitatif vs qualitatif

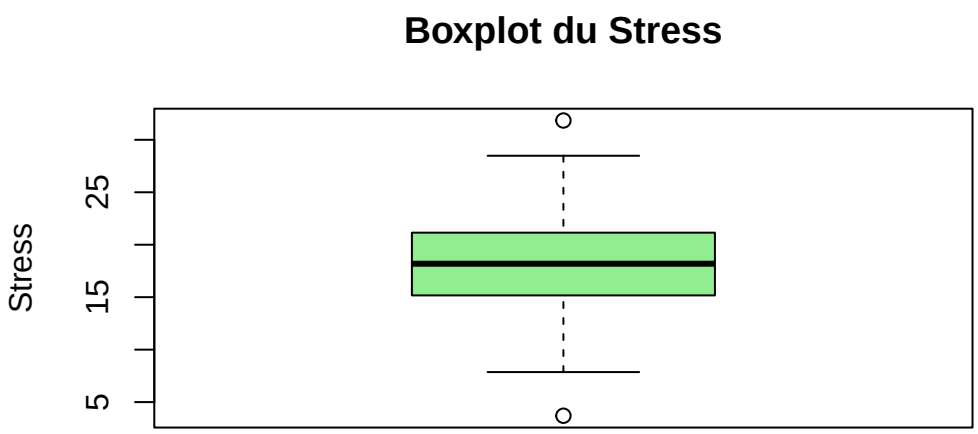
On s'intéresse maintenant au croisement `Stress` vs `EtatCivil`.

On va déterminer s'il existe une relation ou non entre le stress et l'état civil des enseignants interrogés.

a) Donnons le résumé standard de la variable `Stress` :

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.70   15.19   18.19   18.20   21.11   31.84
```

b) Représentons et commentons le boxplot de la variable `Stress` :



- Le boxplot de la variable `Stress` parait symétrique : la médiane se trouvant au milieu et les moustaches ayant a peu près la même longueur avec une valeur extrême de chaque cote.
- Cela veut dire que cette variable suit une loi normale.

c) Constituons 5 classes de mêmes amplitudes pour la variable `Stress`

Table 2.83 – 5 Classes de la variable `Stress`

Stress_classes	Freq
[3.7,9.33]	6
(9.33,15]	34
(15,20.6]	82
(20.6,26.2]	38
(26.2,31.8]	8

d) Donnons les tableaux de contingences de `Stress` vs `EtatCivil` : en effectifs, en fréquences, en pourcentages (arrondir au 100ème près).

Table 2.84 – Tableau de contingence (effectifs) : `Stress` vs `EtatCivil`

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
[3.7,9.33]	2	1	2	1
(9.33,15]	4	3	23	4
(15,20.6]	15	7	58	2
(20.6,26.2]	3	1	34	0
(26.2,31.8]	0	1	7	0

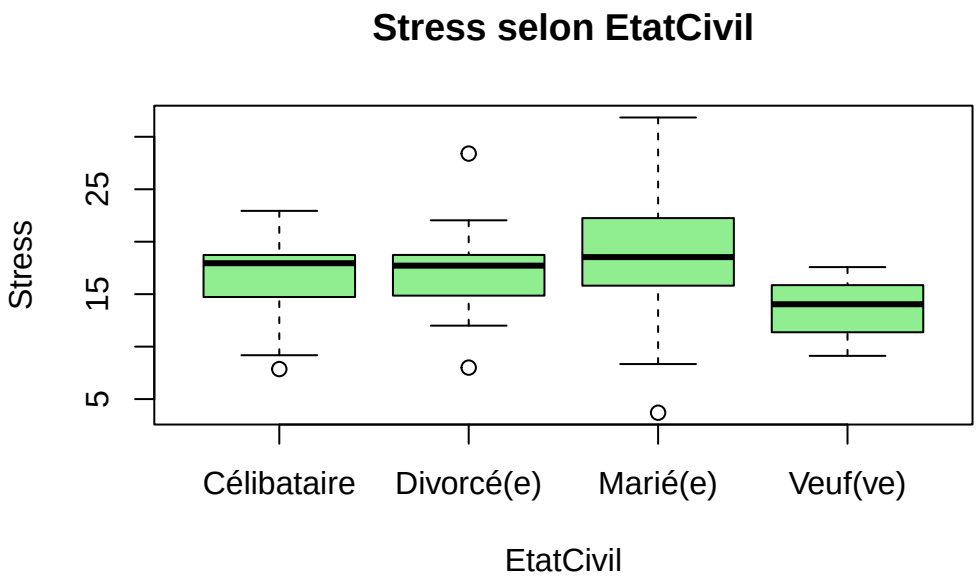
Table 2.85 – Tableau de contingence (fréquences) : `Stress` vs `EtatCivil`

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
[3.7,9.33]	0.01	0.01	0.01	0.01
(9.33,15]	0.02	0.02	0.14	0.02
(15,20.6]	0.09	0.04	0.35	0.01
(20.6,26.2]	0.02	0.01	0.20	0.00
(26.2,31.8]	0.00	0.01	0.04	0.00

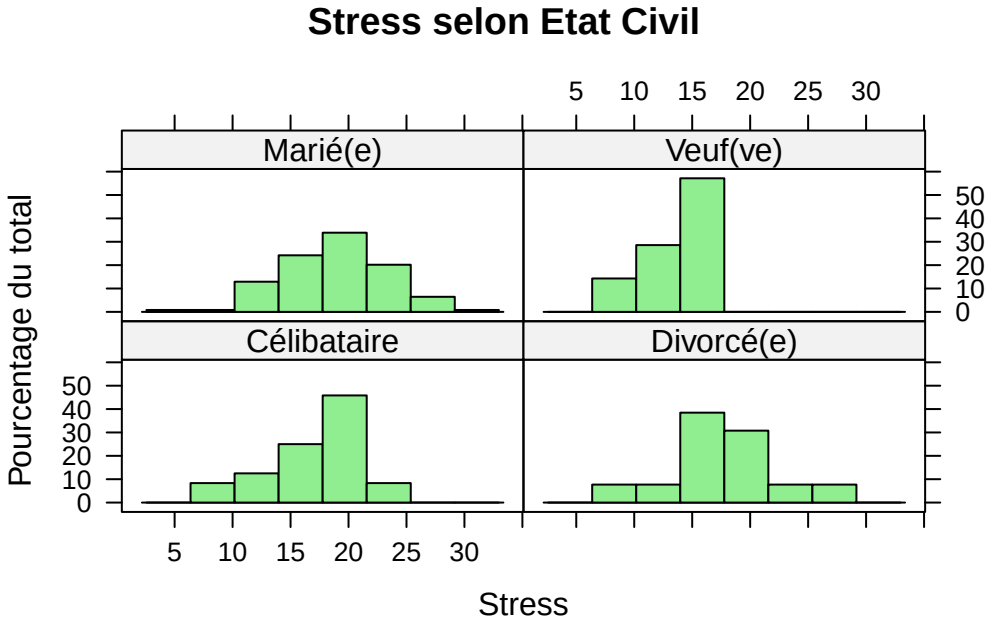
Table 2.86 – Tableau de contingence (pourcentages) : `Stress` vs `EtatCivil`

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)
[3.7,9.33]	1	1	1	1
(9.33,15]	2	2	14	2
(15,20.6]	9	4	35	1
(20.6,26.2]	2	1	20	0
(26.2,31.8]	0	1	4	0

e) Donnons les boxplots de la variable Stress en fonction de la variable EtatCivil. Que remarque-t-on ?



f) Donnons les histogrammes de la variable Stress en fonction de la variable EtatCivil (package lattice)



g) Donnons les résumés de la variable `Stress` en fonction de la variable `EtatCivil` :

```
$Célibataire
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.86   14.91   17.95   16.76   18.66   22.94

$`Divorcé(e)`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   14.86   17.72   17.17   18.74   28.40

$`Marié(e)`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.70   15.82   18.53   18.85   22.25   31.84

$`Veuf(ve)`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.12   11.37   14.04   13.60   15.86   17.58
```

h) Calculons le rapport de corrélation η^2 par étapes (Calcul de la variance intra-groupes, Calcul de la variance inter-groupes, le coefficient η^2) :

```
Rapport de corrélation eta2 calculé manuellement:  0.07492283
```

i) Comparons avec le F de Fisher (Calcul de l'indicateur de Fisher F(`Stress`/`EtatCivil`), Calcul du seuil de signicativité de F(`Stress`/ `EtatCivil`)) :

```
F observée (calculée à partir de eta²) :  4.427502

Seuil de signicativité (F critique) :  2.65972
```

- Puisque la statistique F observée est supérieure a F critique (ou p < 0.05), on **rejette** H0.
- Il existe un écart significatif entre les groupes et donc, il existe une relation statistiquement significative entre `Stress` et `EtatCivil`.

Exercice 22.6. Croisement quantitatif vs quantitatif

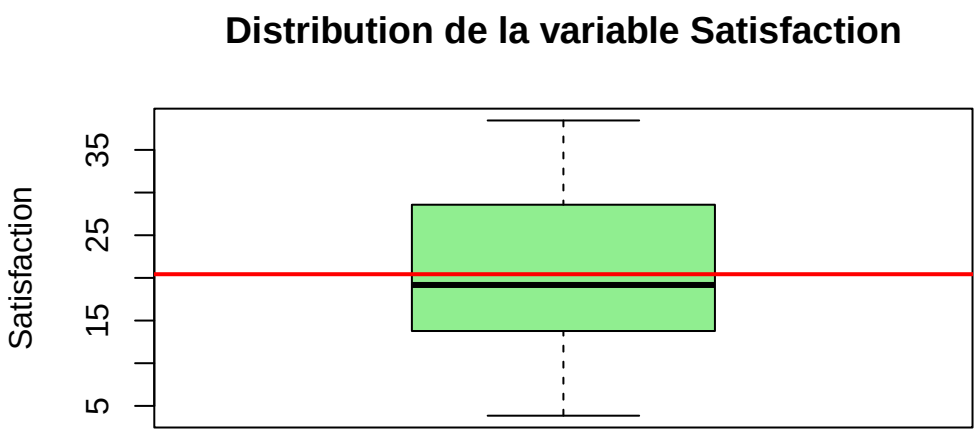
On s'intéresse maintenant au croisement `Age` vs `Satisfaction`.

On va déterminer s'il existe une relation ou non entre l'âge et la satisfaction au travail des enseignants interrogés.

a) Donnons le résumé standard de la variable Satisfaction :

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.85   13.84   19.17   20.43   28.31   38.45
```

b) Représentons et commentons le boxplot de la variable Satisfaction :



c) Détectons graphiquement les "outliers" en dehors d'une bande de 2 écarts-types autour de la moyenne. Pour ce faire, on utilisera la fonction `identify` :

```
[1] 2.016276
```

```
[1] 38.84575
```

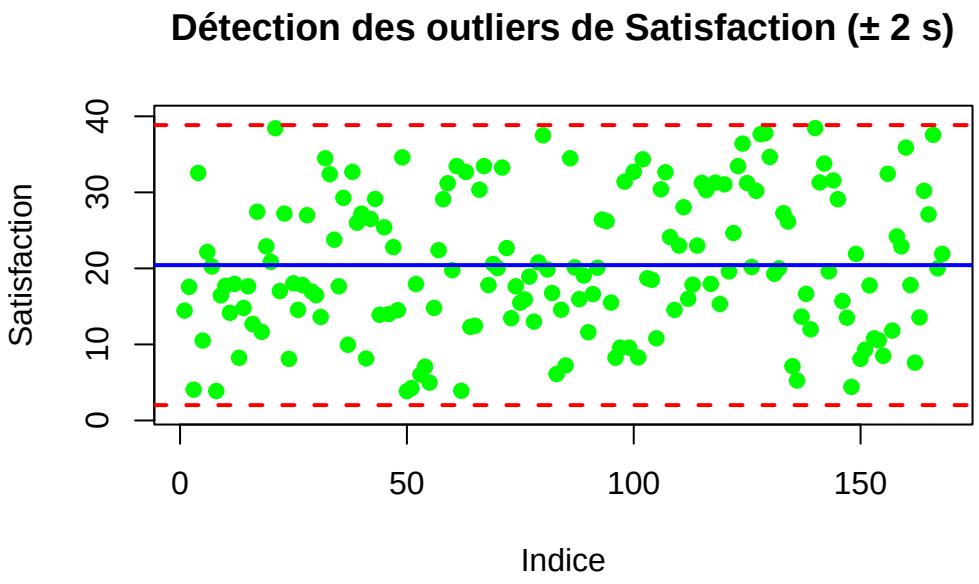


Table 2.87 – Outliers de `Satisfaction` repérés avec `identify`

Sexe	Age	EtatCivil	Nbenfant	Diplome	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	AvisReforme
Homme	56	Veuf(ve)	2	Bac+4	34	1630	38.43	9.12	41.44	Favorable
Homme	53	Marié(e)	1	Bac+4	34	1800	38.45	3.70	42.15	Très favorable

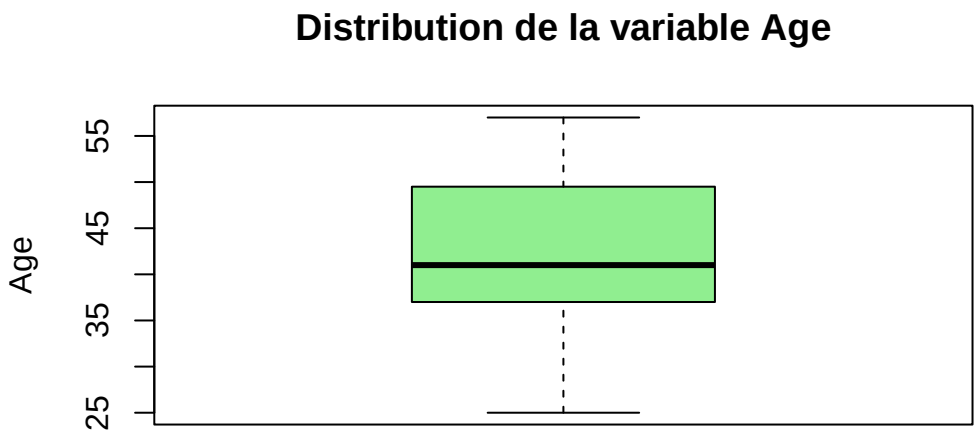
Table 2.88 – Outliers de `Satisfaction` repérés avec les calculs

Sexe	Age	EtatCivil	Nbenfant	Diplome	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	AvisReforme

d) Donnons le résumé standard de la variable `Age` :

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
25.00   37.00   41.00   41.99   49.25   57.00
```

e) Représentons et commentons le boxplot de la variable `Age` :



f) Détectons graphiquement les “outliers” en dehors d’une bande de 2 écarts-types autour de la moyenne :

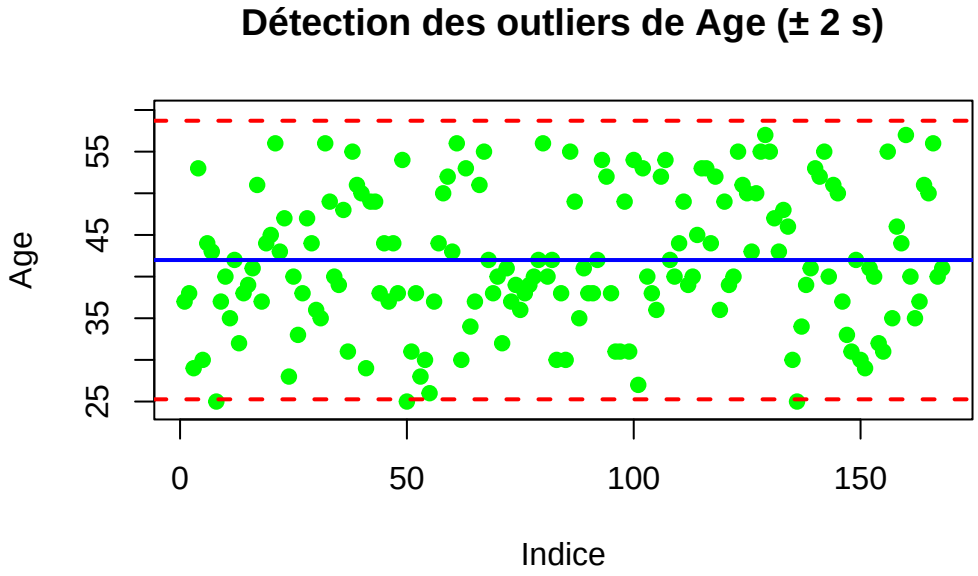


Table 2.89 – Outliers de Age repérés avec identify

Sexe	Age	EtatCivil	Nbenfant	Diplome	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	AvisReforme
Homme	25	Célibataire	0	Bac+6	1	1620	3.87	18.22	3.89	Défavorable
Homme	25	Célibataire	0	Bac+3	1	1600	3.85	18.22	3.89	Très défavorable
Homme	25	Célibataire	0	Bac+3	2	2000	5.25	16.94	5.03	Très favorable

Table 2.90 – Outliers de Age repérés avec les calculs

Sexe	Age	EtatCivil	Nbenfant	Diplome	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	AvisReforme
Homme	25	Célibataire	0	Bac+6	1	1620	3.87	18.22	3.89	Défavorable
Homme	25	Célibataire	0	Bac+3	1	1600	3.85	18.22	3.89	Très défavorable
Homme	25	Célibataire	0	Bac+3	2	2000	5.25	16.94	5.03	Très favorable

g) Constituons 5 classes de même amplitude de la variable Satisfaction :

Table 2.91 – 5 Classes de la variable Satisfaction

classes_Sat	Freq
[3.85,10.8]	27
(10.8,17.7]	43
(17.7,24.6]	42
(24.6,31.5]	30
(31.5,38.5]	26

h) Constituons 4 classes de même amplitude de la variable Age :

Table 2.92 – 4 Classes de la variable Age

classes_Age	Freq
[25,33]	29
(33,41]	60
(41,49]	37
(49,57]	42

i) Donnons les tableaux de contingences de Age vs Satisfaction : en effectifs, en fréquences, en pourcentages (arrondir au 100ème près). Utilisons balloonplot pour voir le tableau de contingence en effectif de Age vs Satisfaction :

Table 2.93 – Tableau de contingence (effectifs) : `Age` vs `Satisfaction`

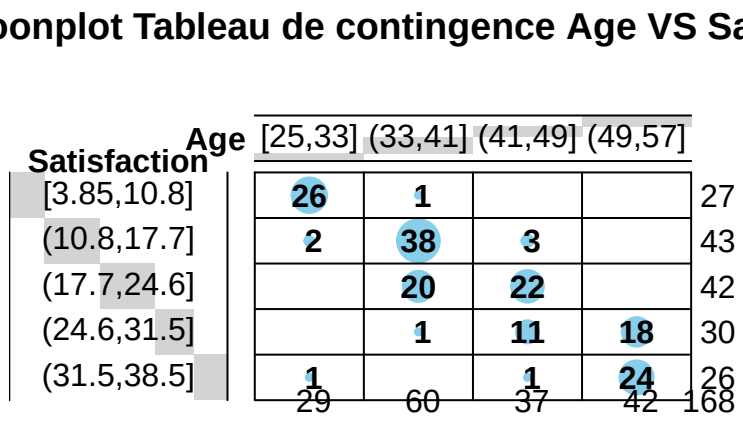
	[3.85,10.8]	(10.8,17.7]	(17.7,24.6]	(24.6,31.5]	(31.5,38.5]
[25,33]	26	2	0	0	1
(33,41]	1	38	20	1	0
(41,49]	0	3	22	11	1
(49,57]	0	0	0	18	24

Table 2.94 – Tableau de contingence (fréquences) : `Age` vs `Satisfaction`

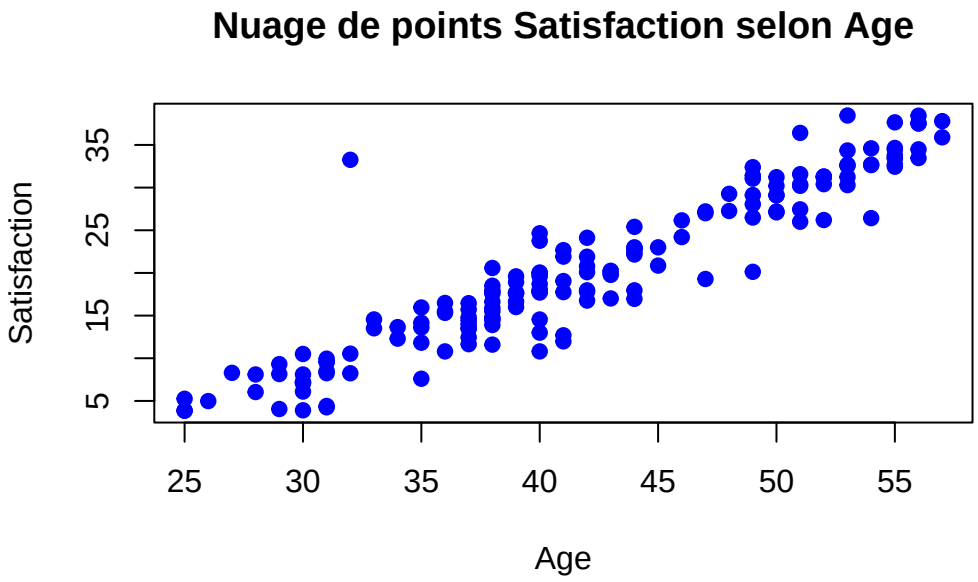
	[3.85,10.8]	(10.8,17.7]	(17.7,24.6]	(24.6,31.5]	(31.5,38.5]
[25,33]	0.155	0.012	0.000	0.000	0.006
(33,41]	0.006	0.226	0.119	0.006	0.000
(41,49]	0.000	0.018	0.131	0.065	0.006
(49,57]	0.000	0.000	0.000	0.107	0.143

Table 2.95 – Tableau de contingence (pourcentages) : `Age` vs `Satisfaction`

	[3.85,10.8]	(10.8,17.7]	(17.7,24.6]	(24.6,31.5]	(31.5,38.5]
[25,33]	15.5	1.2	0.0	0.0	0.6
(33,41]	0.6	22.6	11.9	0.6	0.0
(41,49]	0.0	1.8	13.1	6.5	0.6
(49,57]	0.0	0.0	0.0	10.7	14.3



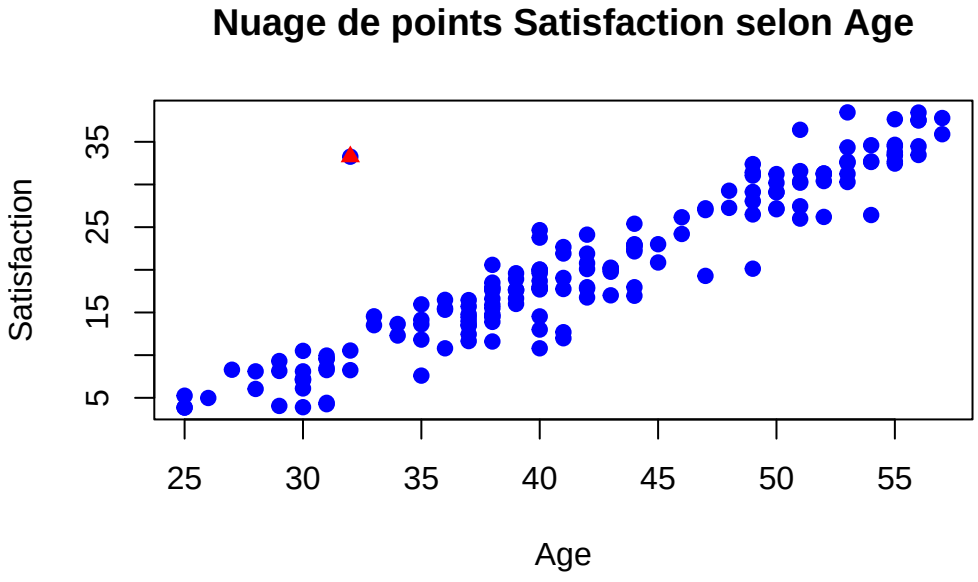
j) Représentons le nuage de points de la variable `Satisfaction` en fonction de la variable `Age` :



k) Détectons dynamiquement un “intrus” avec `identify` et affichons le profil de cet “intrus” :

Table 2.96 – Données de l’intrus

Sexe	Age	EtatCivil	Nbenfant	Diplome	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	AvisReforme
Homme	32	Divorcé(e)	1	Bac+3	30	1660	33.26	8	35.53	Très défavorable

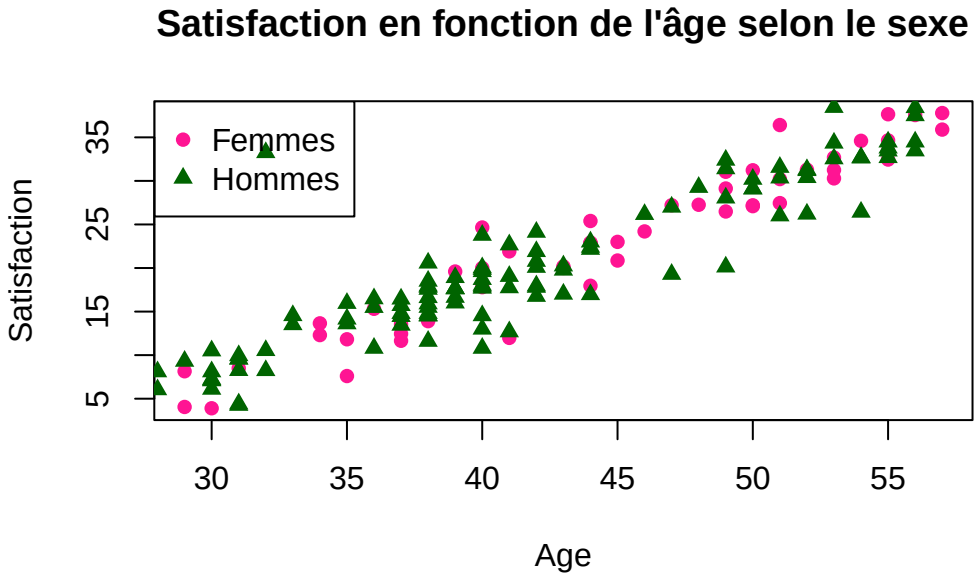


Remarquons que :

- L'individu a 32 ans, ce qui le place en dessous de la moyenne (41,99 ans) et de la médiane de l'échantillon, donc il est relativement jeune.
- Son état civil est divorcé(e) : ce qui est probablement peu fréquent dans l'échantillon.
- Il a un enfant : ce qui est cohérent avec son profil.
- Son niveau de stress est faible (8), tandis que son estime de soi est élevée (35,53), supérieure à la moyenne et à la médiane.

l) Même si cet individu paraît atypique, ses données restent logiques et cohérentes entre elles (Satisfaction, Stress et Estime de soi). Il ne s'agit donc pas d'une erreur de saisie.

m) Représentons le nuage de points de la variable `Satisfaction` en fonction de la variable `Age` en distinguant les femmes et les hommes. On utilisera la fonction `split` :



n) Calculons par étapes la covariance entre la variable `Age` et la variable `Satisfaction`. Comparons avec celle donnée par R :

```
Covariance manuelle:  72.21181

Covariance R:  72.21181
```

o) Calculons par étapes la corrélation entre la variable `Age` et la variable `Satisfaction`. Comparons avec celle donnée par R :

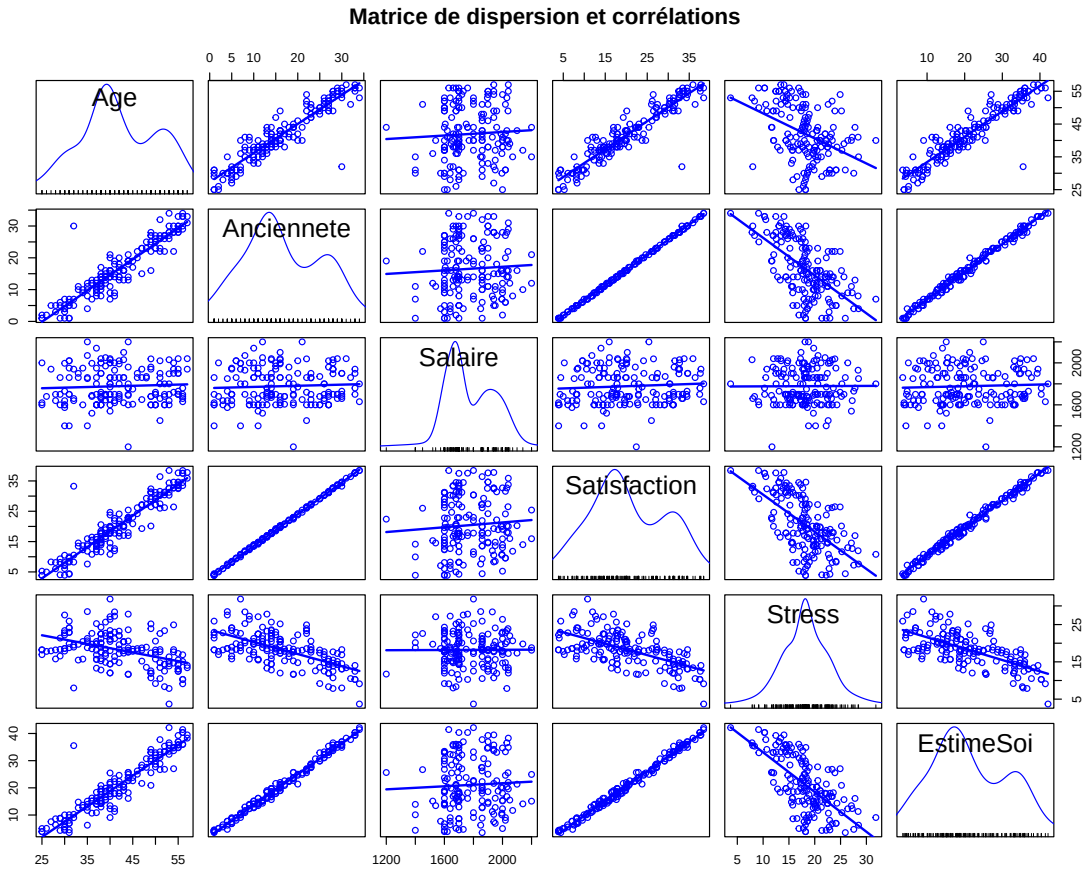
```
Corrélation manuelle:  0.9380404

Corrélation R:  0.9380404
```

p) Donnons la matrice de corrélation entre les variables numériques (sauf le nombre d'enfants) :

Table 2.97 – Matrice de corrélation sauf `Nbenfant`

	Age	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi
Age	1.0000000	0.9323131	0.0559726	0.9380404	-0.4200334	0.9354238
Anciennete	0.9323131	1.0000000	0.0553876	0.9996856	-0.6200661	0.9952135
Salaire	0.0559726	0.0553876	1.0000000	0.0742909	0.0055120	0.0485073
Satisfaction	0.9380404	0.9996856	0.0742909	1.0000000	-0.6119429	0.9951199
Stress	-0.4200334	-0.6200661	0.0055120	-0.6119429	1.0000000	-0.6664936
EstimeSoi	0.9354238	0.9952135	0.0485073	0.9951199	-0.6664936	1.0000000

q) Représentons cette matrice graphiquement et commentons. On pourra utiliser la fonction `scatterplotMatrix` :



On peut remarquer que :

- Les variables `Age` , `Satisfaction` et `Anciennete` sont fortement corrélées positivement entre elles.
- Par ailleurs, la variable `Satisfaction` est aussi fortement corrélée positivement avec la variable `EstimeSoi`.
- En outre, on peut voir que les distributions des variables `Age` , `Satisfaction`, `Anciennete` et `EstimeSoi` ont à peu près la même allure.

- La distribution de la variable Salaire a la même queue à la fin mais diffère complètement au début.
- La distribution de la variable `Stress` est, quant à elle, symétrique et ressemble à une distribution normale.

3 Travail personnel 3 (ACP)

Exercice 23.1 Création du jeu de données X

Créons le jeu de données X :

Table 3.1 – Jeu de données X

Z1	Z2
1	5
2	10
3	8
4	8
9	12

Table 3.2 – Jeu de données X centrées réduites

Z1	Z2
-0.8990258	-1.3805370
-0.5779452	0.5368755
-0.2568645	-0.2300895
0.0642161	-0.2300895
1.6696194	1.3038405

Table 3.3 – Valeurs propres

x
1.7880244
0.2119756

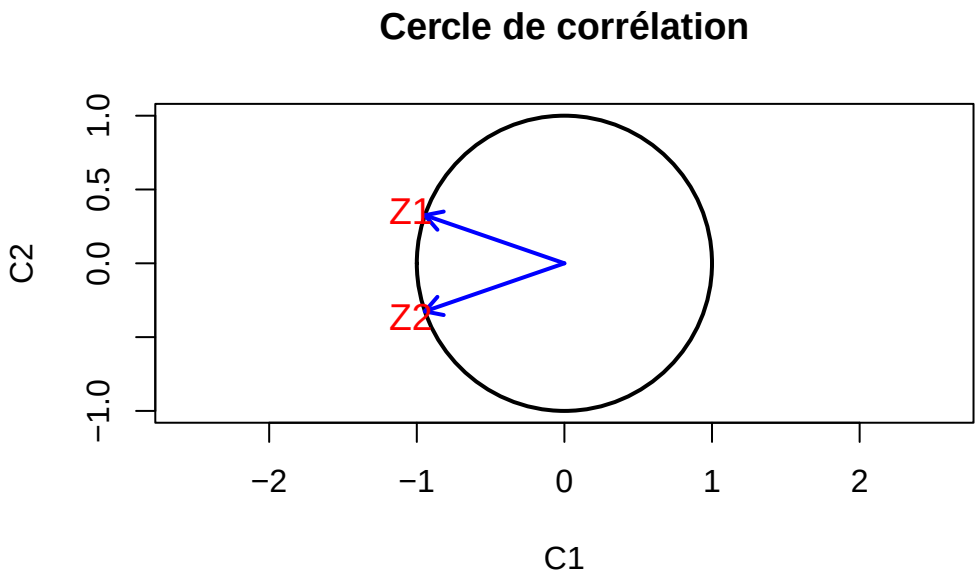
Table 3.4 – Vecteurs propres

-0.7071068	0.7071068
-0.7071068	-0.7071068

```
Le pourcentage de variance expliquée par le 1er axe est:  89.4

Le pourcentage de variance expliquée par le 2eme axe est:  10.6
```

Table 3.5 – Coordonnées des variables

	C1	C2
Z1	-0.9455222	0.3255577
Z2	-0.9455222	-0.3255577




```
Coefficient de corrélation entre X1 et X2:  0.7880244
```

Table 3.6 – Coordonnées des individus

1.6118943	0.3404798
0.0290406	-0.7882972
0.3443285	-0.0189328
0.1172902	0.2081055
-2.1025536	0.2586447

Exercice 23.3 Comparaison des résultats de `princomp` et `prcomp`

```
[1] 1.7880244 0.2119756
```

```
   Comp.1    Comp.2
1.7880244 0.2119756
```

Table 3.7 – Valeurs propres `prcomp`

x
1.7880244
0.2119756

Table 3.8 – Valeurs propres `princomp`

	x
Comp.1	1.7880244
Comp.2	0.2119756

Table 3.9 – Vecteurs propres `prcomp`

	PC1	PC2
Z1	0.7071068	0.7071068
Z2	0.7071068	-0.7071068

Table 3.10 – Vecteurs propres `princomp`

	Comp.1	Comp.2
Z1	0.7071068	0.7071068
Z2	0.7071068	-0.7071068

Table 3.11 – Coordonnées des individus `prcomp`

PC1	PC2
-1.6118943	0.3404798
-0.0290406	-0.7882972
-0.3443285	-0.0189328
-0.1172902	0.2081055
2.1025536	0.2586447

Table 3.12 – Coordonnées des individus `princomp`

Comp.1	Comp.2
-1.8021526	0.3806680
-0.0324684	-0.8813431
-0.3849710	-0.0211675
-0.1311344	0.2326690
2.3507264	0.2891736

Table 3.13 – Valeurs propres PCA

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	1.7880244	89.40122	89.40122
comp 2	0.2119756	10.59878	100.00000

Table 3.14 – Coordonnées des individus PCA

Dim.1	Dim.2
-1.8021526	0.3806680
-0.0324684	-0.8813431
-0.3849710	-0.0211675
-0.1311344	0.2326690
2.3507264	0.2891736

Exercice 24. Données `stations`

Chargeons les données, regardons la structure et affichons les premieres lignes :

Table 3.15 – Structure des données

Variable	Type	Valeurs
Nom	chr	: chr ”LesAillons” ”LesArcs” ”Arèches” ”Aussois” ...
prixforf	int	76 160 85 71 54 79 88 140 82 42 ...
altmin	int	: int 900 800 750 500 1710 1850 1550 1100 1230 1000 ...
altmax	int	: int 2000 3226 2300 2750 2200 3000 1850 2707 1650 1600 ...
pistes	int	: int 45 117 30 21 4 16 36 100 26 0 ...
kmfond	int	: int 50 30 47 10 80 0 25 66 7 12 ...
remontee	int	22 69 15 11 4 10 24 67 17 40 ...

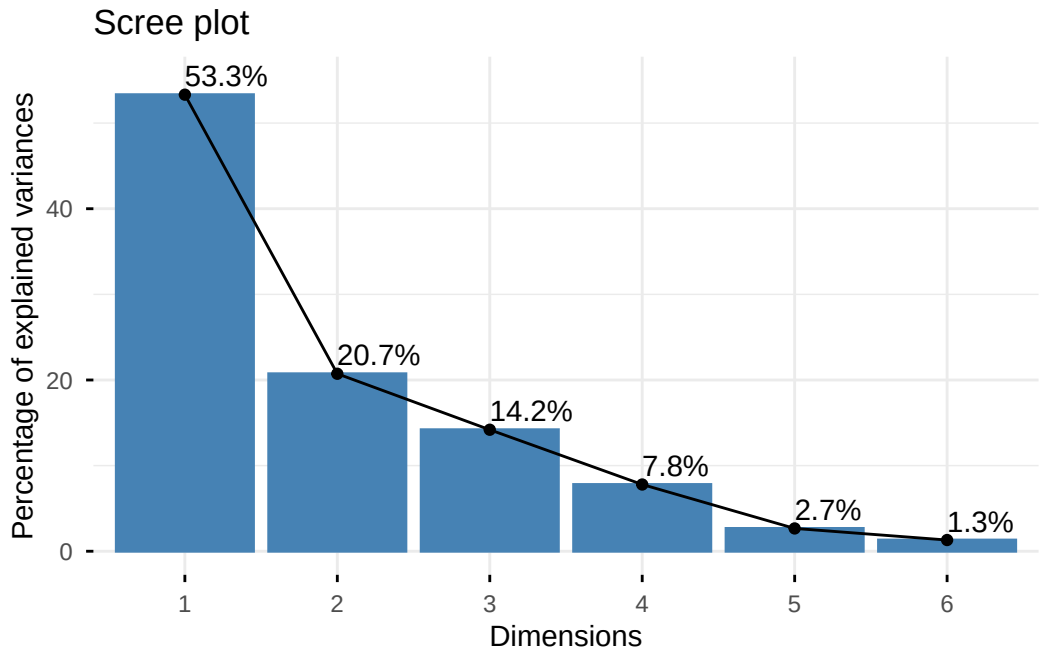
Table 3.16 – Aperçu des données `stations`

Nom	prixforf	altmin	altmax	pistes	kmfond	remontee
LesAillons	76	900	2000	45	50	22
LesArcs	160	800	3226	117	30	69
Arèches	85	750	2300	30	47	15
Aussois	71	500	2750	21	10	11
Bessans	54	1710	2200	4	80	4
Bonneval	79	1850	3000	16	0	10

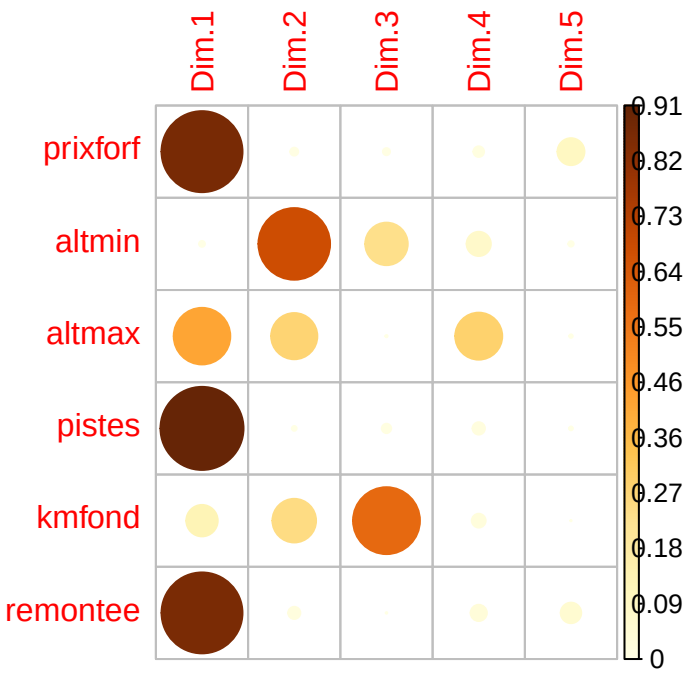
Nos données sont quantitatives, sauf pour la variable `Nom`. On l’excluera pour réaliser une ACP sur le jeu de données. Affichons les valeurs propres et visualisons le screeplot :

Table 3.17 – Composantes Principales `PCA` : Données `stations`

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.1990602	53.317670	53.31767
comp 2	1.2435500	20.725833	74.04350
comp 3	0.8513896	14.189827	88.23333
comp 4	0.4676520	7.794200	96.02753
comp 5	0.1597775	2.662958	98.69049
comp 6	0.0785707	1.309512	100.00000



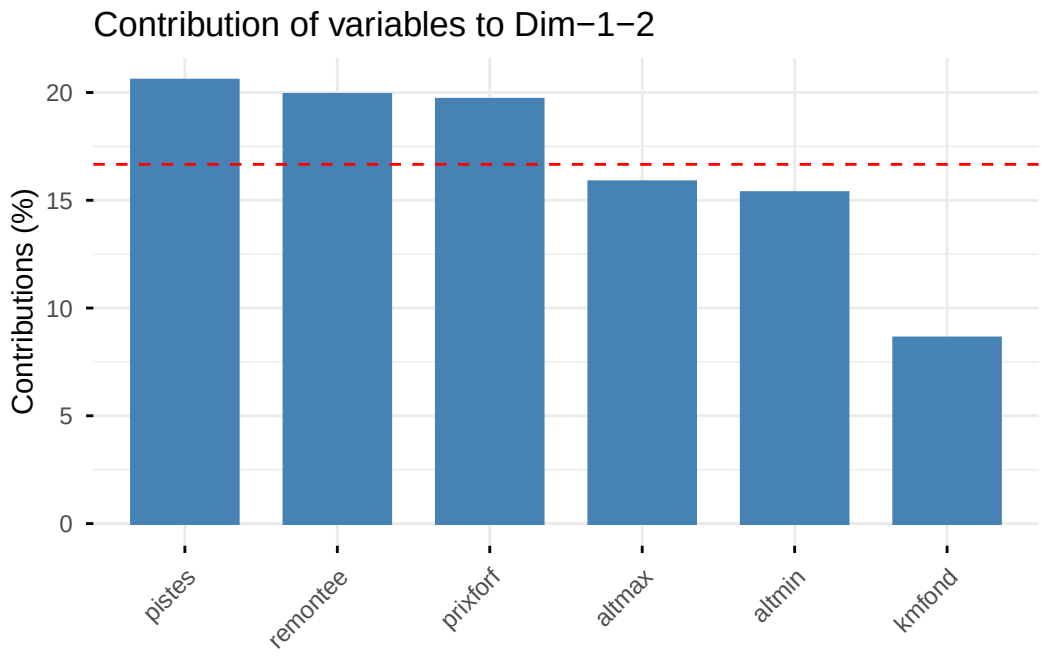
Visualisons maintenant la qualité de représentation des variables avec `corrplot` :

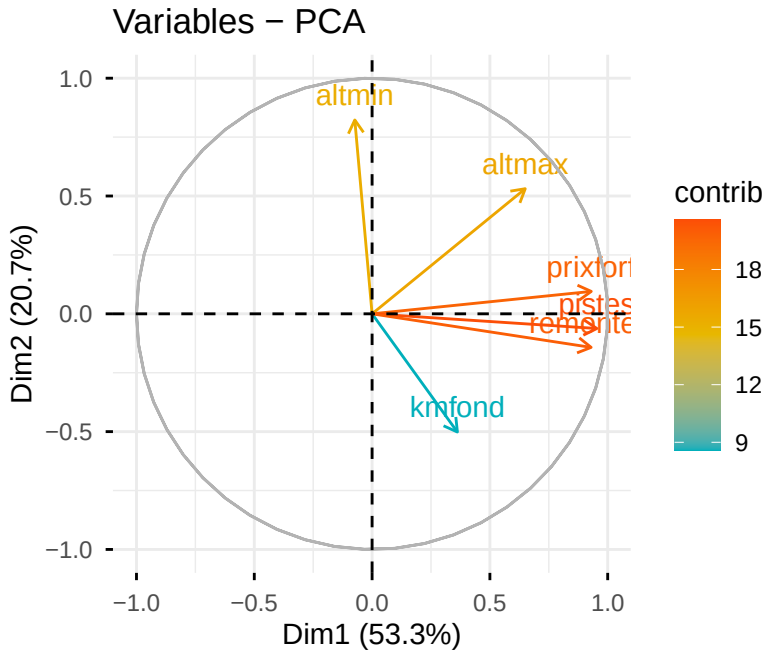


- On remarque que les variables sont `prixforf`, `pistes` et `remontee` sont tres bien représentées sur la premiere composante. Par ailleurs, la variable `altmin` est bien représentée sur la deuxieme composante et `kmfond` sur la 3 eme.
- La variable est la moins expliquée par les 3 premieres dimensions.

Visualisons maintenant la contribution des variables aux 2 premières composantes principales :

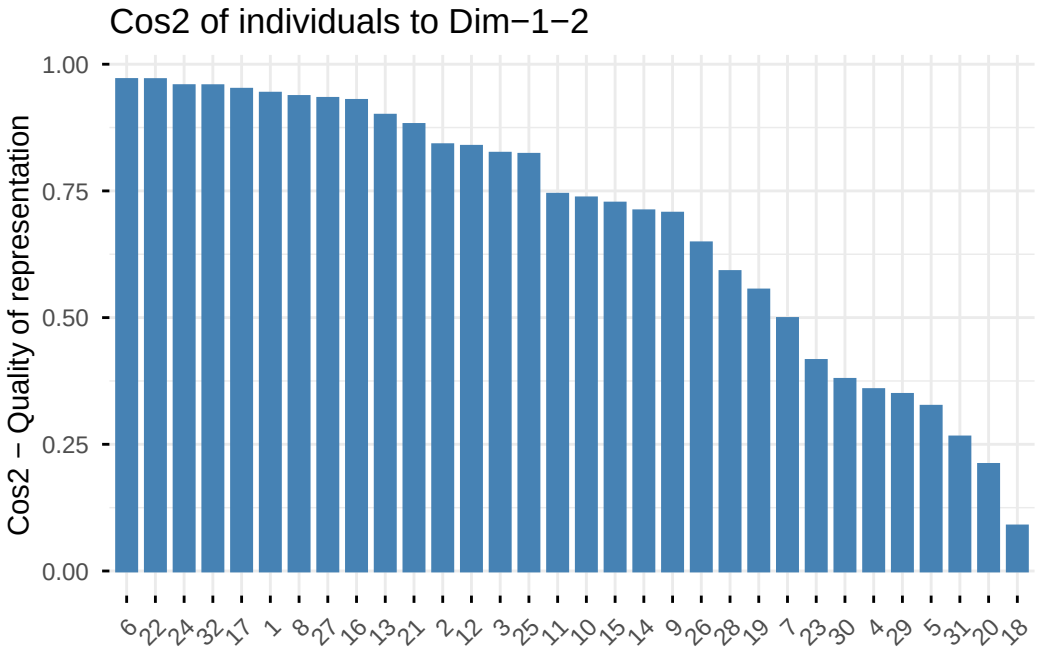
```
              Dim.1      Dim.2      Dim.3     Dim.4      Dim.5
prixforf 27.0545026  0.7277779  0.8630735  3.348903 60.3619164
altmin    0.1682591 54.4283218 28.0528130 16.673650  0.5278217
altmax   13.2095337 22.6732235  0.1724779 62.326754  1.5443938
pistes   28.4521266  0.3117897  1.4427227  4.472173  1.7256503
kmfond    4.0948176 20.2283699 69.3313170  5.564903  0.6614036
remontee 27.0207604  1.6305172  0.1375958  7.613617 35.1788143
```

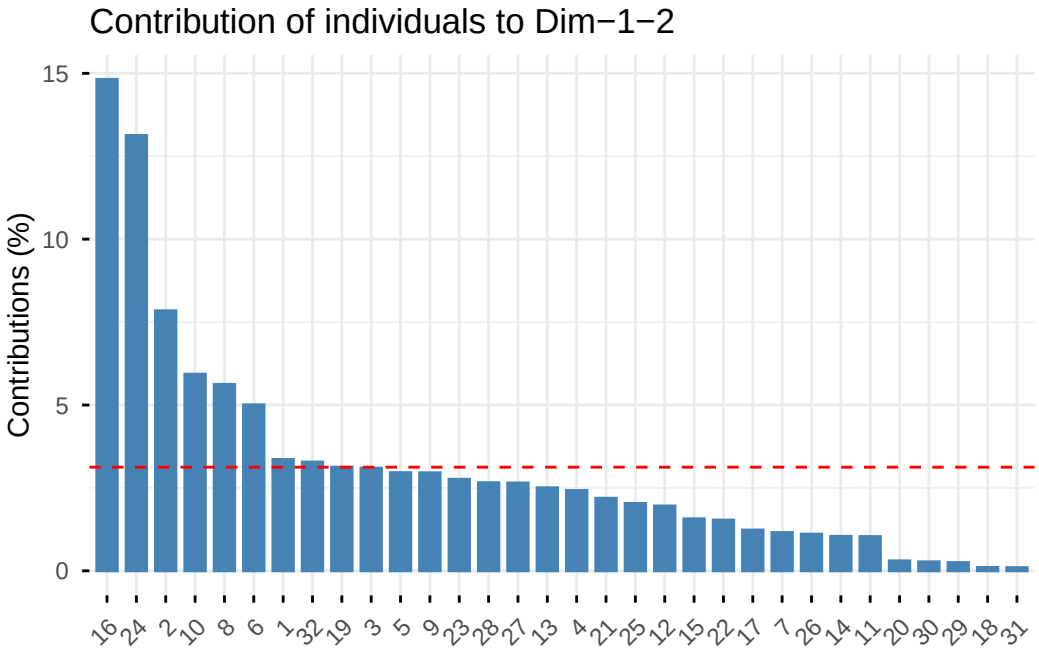


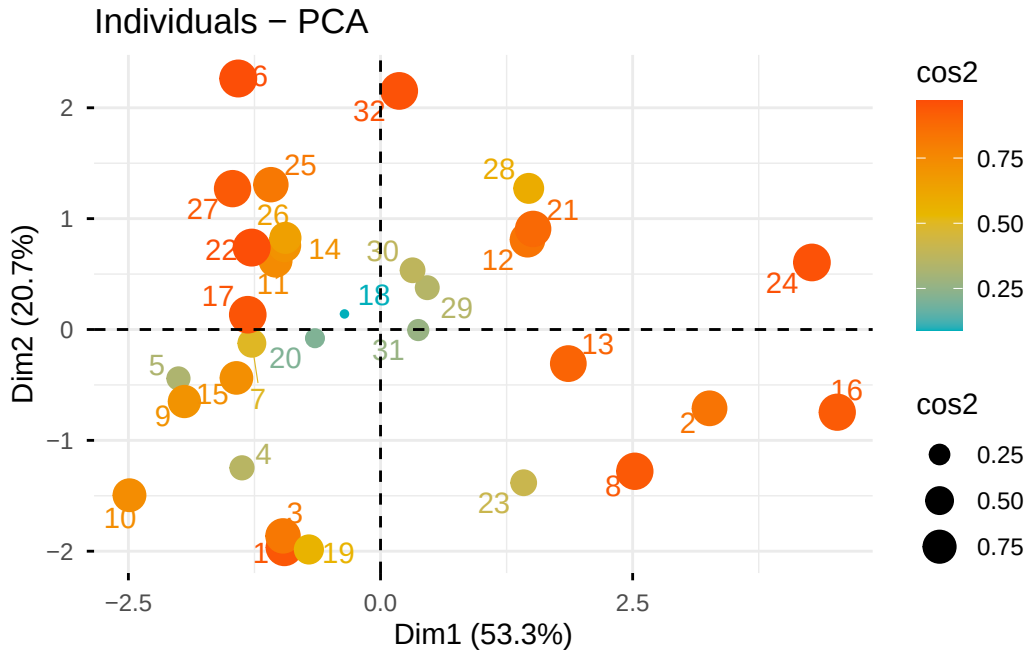


- Remarquons que les variables les plus contributives aux dimensions 1 et 2 sont `pistes`, `remontee` et `prixforf`.

Visualisons maintenant la qualité de représentation et la contribution des individus aux 2 premières composantes principales :


Cos2 of individuals to Dim−1−2
Cos2 – Quality of representation
1.00
0.75
0.50
0.25
0.00
6 22 24 32 17 1 8 27 16 13 21 2 12 3 25 11 10 15 14 9 26 28 19 7 23 30 4 29 5 31 20 18


Contribution of individuals to Dim−1−2
Contributions (%)
15
10
5
0
16 24 2 10 8 6 1 32 19 3 5 9 23 28 27 13 4 21 25 12 15 22 17 7 26 14 11 20 30 29 18 31



- Les individus les mieux représentés par les 2 premières composantes principales sont ceux avec un cos2 supérieur a 0.75.
- Les individus les plus contributifs (contribuent plus que la moyenne) aux 2 premières composantes principales sont : 16, 24, 2, 10, 8, 6, 1 et 32.

4 Travail personnel 4 (AFC)

Exercice 31.1 Données `USArrests`

Le jeu de données `USArrests` a été chargé et avec la fonction `class`, on obtient facilement sa classe. Affichons les premières lignes :

```
La classe de cet objet est:  data.frame
```

Table 4.1 – Aperçu des données `USArrests`

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

Exercice 31.2. Fonctions `princomp` et `prcomp`

Avec les fonctions `princomp` et `prcomp`, nous calculons les composantes principales de l'ACP (les scores des 50 états). Ici, on affiche que les scores des 5 premiers états selon chaque méthode :

Table 4.2 – Aperçu des scores des 50 états (`prcomp`)

	PC1	PC2	PC3	PC4
Alabama	-0.9756604	-1.1220012	0.4398037	0.1546966
Alaska	-1.9305379	-1.0624269	-2.0195003	-0.4341755
Arizona	-1.7454429	0.7384595	-0.0542302	-0.8262642
Arkansas	0.1399989	-1.1085423	-0.1134222	-0.1809736
California	-2.4986128	1.5274267	-0.5925410	-0.3385592
Colorado	-1.4993407	0.9776297	-1.0840016	0.0014502

Table 4.3 – Aperçu des scores des 50 états (`princomp`)

	Comp.1	Comp.2	Comp.3	Comp.4
Alabama	0.9855659	1.1333924	0.4442688	0.1562671
Alaska	1.9501378	1.0732133	-2.0400033	-0.4385834
Arizona	1.7631635	-0.7459568	-0.0547808	-0.8346529
Arkansas	-0.1414203	1.1197968	-0.1145737	-0.1828109
California	2.5239801	-1.5429340	-0.5985568	-0.3419965
Colorado	1.5145629	-0.9875551	-1.0950070	0.0014649

- On observe de petites différences numériques entre les résultats fournis par `prcomp` et `princomp`. Les autres différences observées portent uniquement sur le signe des axes principaux.

Exercice 31.3 Composantes principales avec notre fonction `gsvd`

1. Après standardisation des données `USArrests` avec la fonction `scale`, nous obtenons :

Table 4.4 – Aperçu des données `USArrests` standardisées

	Murder	Assault	UrbanPop	Rape
Alabama	1.2425641	0.7828393	-0.5209066	-0.0034165
Alaska	0.5078625	1.1068225	-1.2117642	2.4842029
Arizona	0.0716334	1.4788032	0.9989801	1.0428784
Arkansas	0.2323494	0.2308680	-1.0735927	-0.1849166
California	0.2782682	1.2628144	1.7589234	2.0678203
Colorado	0.0257146	0.3988593	0.8608085	1.8649672

2. Pour calculer les composantes principales de l'ACP avec la fonction manuelle `gsvd`, nous avons d'abord créé la dite fonction et en l'appliquant à notre jeu de données, nous avons obtenu les scores ci-dessous pour les 5 premiers états :

Table 4.5 – Aperçu des scores des 50 états (`gsvd`)

Alabama	-0.9756604	-1.1220012	0.4398037	0.1546966
Alaska	-1.9305379	-1.0624269	-2.0195003	-0.4341755
Arizona	-1.7454429	0.7384595	-0.0542302	-0.8262642

Arkansas	0.1399989	-1.1085423	-0.1134222	-0.1809736
California	-2.4986128	1.5274267	-0.5925410	-0.3385592
Colorado	-1.4993407	0.9776297	-1.0840016	0.0014502

3. Si on compare les résultats trouvés avec `gsvd` à ceux de `princomp` et `prcomp`, on peut voir que les valeurs des scores sont très proches, avec parfois un changement de signe sur certains axes.

Table 4.6 – Aperçu des scores des 50 états (`PCA`)

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.9855659	-1.1333924	0.4442688	0.1562671
Alaska	1.9501378	-1.0732133	-2.0400033	-0.4385834
Arizona	1.7631635	0.7459568	-0.0547808	-0.8346529
Arkansas	-0.1414203	-1.1197968	-0.1145737	-0.1828109
California	2.5239801	1.5429340	-0.5985568	-0.3419965
Colorado	1.5145629	0.9875551	-1.0950070	0.0014649

4. Les résultats trouvés avec `gsvd` et ceux de la fonction `PCA` du package `FactoMineR` sont aussi très proches. Les changements de signe sur certains axes sont visibles mais cela ne change pas l'interprétation.

Exercice 32. Etude du lien entre les variables `CSP` et `HEB`

On considère un ensemble de `18282` individus pour lesquels on connaît la CSP (modalités agriculteur AGRI, cadre supérieur `CADR`, inactif `INAC`, et ouvrier `OUVR`) et le choix d'hébergement pour les vacances HEB (modalités camping `CAMP`, `HOTEL`, location `LOCA`, et résidence secondaire `RESI`).

Dans cet exercice, notre but sera de représenter les éventuels liens entre la catégorie socio-professionnelle `CSP` et le type d'hébergement choisi `HEB`.

1. Créons le tableau de contingence et calculons la statistique du khi-deux avec `chisq.test`

Table 4.7 – Tableau de contingence : `CSP` & `HEB`

	CAMP	HOTEL	LOCA	RESI
AGRI	239	155	129	0
CADR	1003	1556	1821	1521
INAC	682	1944	967	1333
OUVR	2594	1124	2176	1038

```
    Pearson's Chi-squared test

data:  tab_CSP_HEB
X-squared = 2067.9, df = 9, p-value < 2.2e-16
```

- Selon le résultat du test de χ^2, la statistique vaut `2067.9` et la p-valeur est inférieure à `0.05`. On **rejette** donc l'hypothese d'indépendance des 2 variables `CSP` et `HEB`.
- Elles sont statistiquement significativement associées.

Par conséquent, on peut réaliser une AFC.

2. Donnons les profils-lignes et les profils-colonnes :

Table 4.8 – Profils-lignes : `CSP` & `HEB`

	CAMP	HOTEL	LOCA	RESI
AGRI	0.4569790	0.2963671	0.2466539	0.0000000
CADR	0.1699712	0.2636841	0.3085918	0.2577529
INAC	0.1384490	0.3946407	0.1963053	0.2706050
OUVR	0.3742066	0.1621466	0.3139065	0.1497403

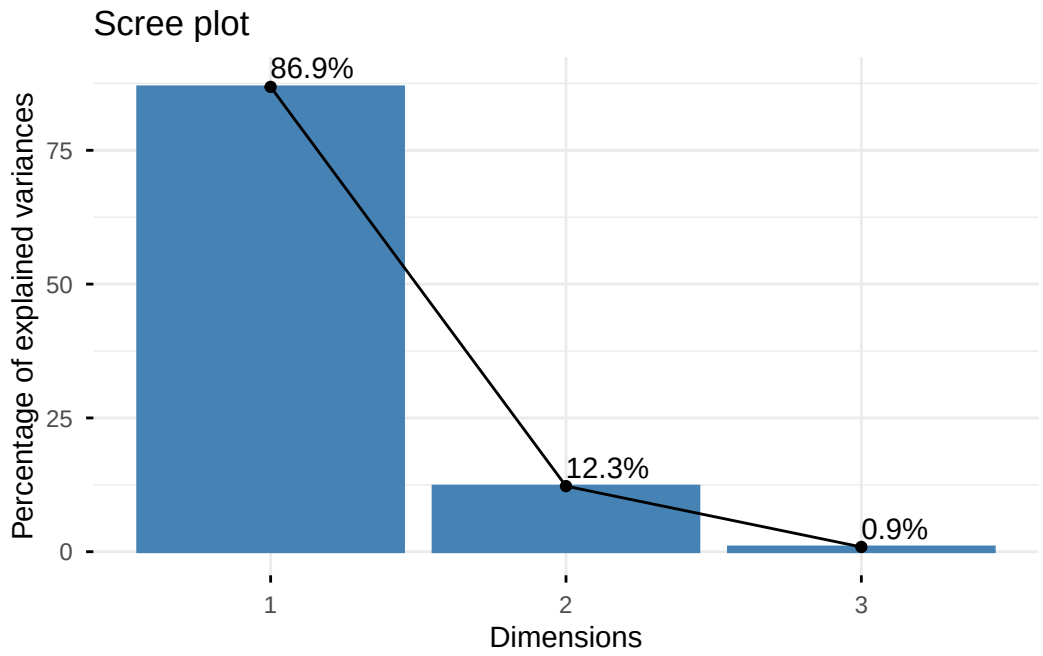
Table 4.9 – Profils-colonnes : `CSP` & `HEB`

	CAMP	HOTEL	LOCA	RESI
AGRI	0.0528995	0.0324336	0.0253289	0.0000000
CADR	0.2220009	0.3255911	0.3575496	0.3908016
INAC	0.1509517	0.4067797	0.1898684	0.3424974
OUVR	0.5741479	0.2351956	0.4272531	0.2667009

On observe que :

- 45% des agriculteurs choisissent `camping` comme hébergement.
- 57% des campings sont habités par des ouvriers.
- Les agriculteurs sont très rares dans tous les types d'hébergement.

3. Réalisons une AFC sur le tableau de contingence :



- La première dimension explique, à elle seule, 86,9 % de l'inertie totale. Néanmoins, afin de représenter les relations entre modalités dans un plan factoriel, il est d'usage de retenir les deux premiers axes.

4. Vérifions que la statistique de χ^2 égale la somme des valeurs propres multipliée par n :

```
Chi2: 2067.911
```

```
n * somme des valeurs propres: 2067.911
```

5. Donnons les coordonnées des modalités :

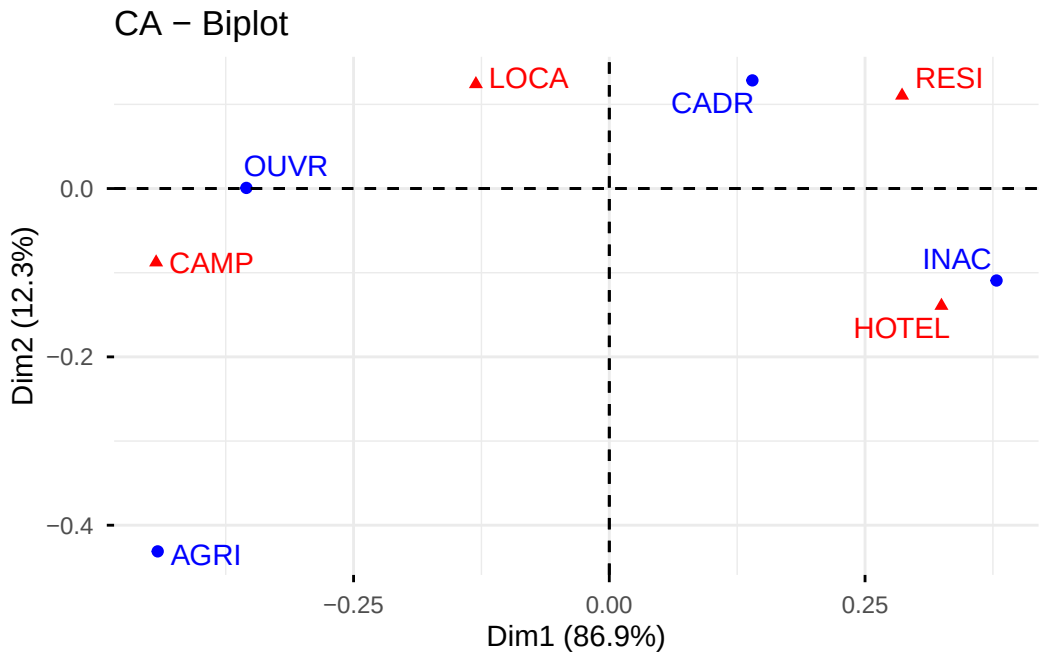
Table 4.10 – Coordonnées des modalités lignes

	Dim 1	Dim 2	Dim 3
AGRI	-0.4414992	-0.4310783	-0.1366310
CADR	0.1399003	0.1285091	-0.0266734
INAC	0.3785766	-0.1092544	0.0198391
OUVR	-0.3548061	0.0007658	0.0189166

Table 4.11 – Coordonnées des modalités colonnes

	Dim 1	Dim 2	Dim 3
CAMP	-0.4430269	-0.0877102	0.0222801
HOTEL	0.3247191	-0.1393080	-0.0188210
LOCA	-0.1304042	0.1241100	-0.0362286
RESI	0.2862054	0.1104664	0.0446549

Représentons les graphiquement avec un biplot :



Analysons un peu le graphe :

- On peut vite remarquer que les ouvriers sont le plus associés aux campings.
- Les cadres sont le plus associés aux résidences ; et les inactifs aux hotels.
- Par contre, les agriculteurs ne sont associés à aucun type d'hébergement.

6. Donnons les contributions et cosinus carrés :

Table 4.12 – Contributions des modalités lignes

	Dim 1	Dim 2	Dim 3
AGRI	5.675896	38.3469998	53.11637
CADR	6.430372	38.4512978	22.84068
INAC	39.307445	23.2000982	10.54792
OUVR	48.586287	0.0016042	13.49503

Table 4.13 – Cosinus carrés des modalités lignes

	Dim 1	Dim 2	Dim 3
AGRI	0.4880138	0.4652482	0.0467381
CADR	0.5318771	0.4487886	0.0193343
INAC	0.9207832	0.0766881	0.0025287
OUVR	0.9971609	0.0000046	0.0028345

Table 4.14 – Contributions des modalités colonnes

	Dim 1	Dim 2	Dim 3
CAMP	49.37184	13.71399	12.201331
HOTEL	28.05599	36.59369	9.209846
LOCA	4.82203	30.95314	36.366826
RESI	17.75013	18.73917	42.221997

Table 4.15 – Cosinus carrés des modalités colonnes

	Dim 1	Dim 2	Dim 3
CAMP	0.9599462	0.0376259	0.0024278
HOTEL	0.8421693	0.1550014	0.0028292
LOCA	0.5042918	0.4567855	0.0389226
RESI	0.8522858	0.1269667	0.0207476

7. Oui, il y a un effet Guttman :

- La première dimension de l'AFC explique la majorité de l'inertie.
- Les autres dimensions apportent très peu d'information.
- Presque toutes les modalités lignes et colonnes sont bien représentées sur Dim 1.

Exercice 33. Données `smoke`

1. Chargeons le jeu de données `smoke` du package `ca` dans R avec la commande `data`. Affichons les premières données :

Table 4.16 – Aperçu des données `smoke`

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

2. **AFC et SVD généralisée**

a) Construisons la matrice `F` des fréquences :

Table 4.17 – Matrice `F` des fréquences

	none	light	medium	heavy
SM	0.0207254	0.0103627	0.0155440	0.0103627
JM	0.0207254	0.0155440	0.0362694	0.0207254
SE	0.1295337	0.0518135	0.0621762	0.0207254
JE	0.0932642	0.1243523	0.1709845	0.0673575
SC	0.0518135	0.0310881	0.0362694	0.0103627

Vecteurs r et c des distributions marginales :

Table 4.18 – Distribution marginale ligne (`smoke`)

	x
SM	0.0569948
JM	0.0932642
SE	0.2642487
JE	0.4559585
SC	0.1295337

Table 4.19 – Distribution marginale colonne (`smoke`)

	x
none	0.3160622
light	0.2331606
medium	0.3212435
heavy	0.1295337

Matrice `Z` des écarts à l'indépendance :

```
           none         light        medium        heavy
SM  0.002711482 -0.0029262531 -0.002765175  0.002979946
JM -0.008751913 -0.0062015088  0.006308894  0.008644527
SE  0.046014658 -0.0097989208 -0.022712019 -0.013503718
JE -0.050847003  0.0180407528  0.024510725  0.008295525
SC  0.010872775  0.0008859298 -0.005342425 -0.006416280
```

Table 4.20 – Matrice `Z` des écarts à l'indépendance

	none	light	medium	heavy
SM	0.0027115	-0.0029263	-0.0027652	0.0029799
JM	-0.0087519	-0.0062015	0.0063089	0.0086445
SE	0.0460147	-0.0097989	-0.0227120	-0.0135037
JE	-0.0508470	0.0180408	0.0245107	0.0082955
SC	0.0108728	0.0008859	-0.0053424	-0.0064163

b) Calculer avec la fonction `gsvd` les matrices X et Y des coordonnées factorielles des profil-lignes et colonnes de l'AFC.

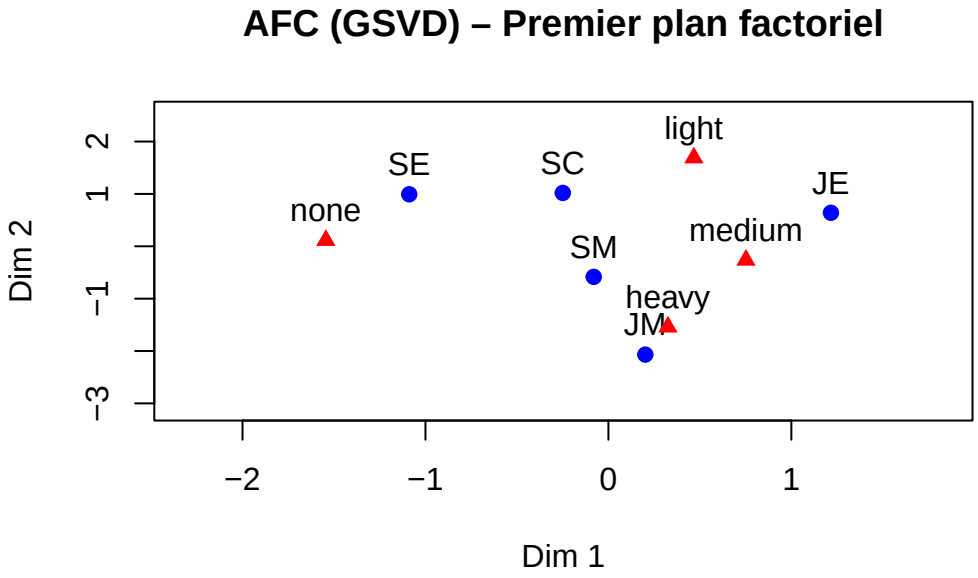
Table 4.21 – Coordonnées factorielles des profil-lignes (`smoke`)

	Dim1	Dim2	Dim3
SM	-0.0803894	-0.5836613	3.1404838
JM	0.2018368	-2.0681186	-1.8074457
SE	-1.0886327	0.9930016	-0.3484710
JE	1.2163753	0.6397322	-0.1424654
SC	-0.2491899	1.0190460	-0.8421018

Table 4.22 – Coordonnées factorielles des profil-colonnes (`smoke`)

	Dim1	Dim2	Dim3
none	-1.5443458	0.1136213	-0.4675176
light	0.4670906	1.6904299	0.4552029
medium	0.7518820	-0.2638305	-1.3957219
heavy	0.3253732	-1.5402208	1.4080366

c) Représentons avec la fonction `plot` les profil-lignes et les profil-colonnes sur le premier plan factoriel de l'AFC.



Sur ce premier plan factoriel, on peut observer que :

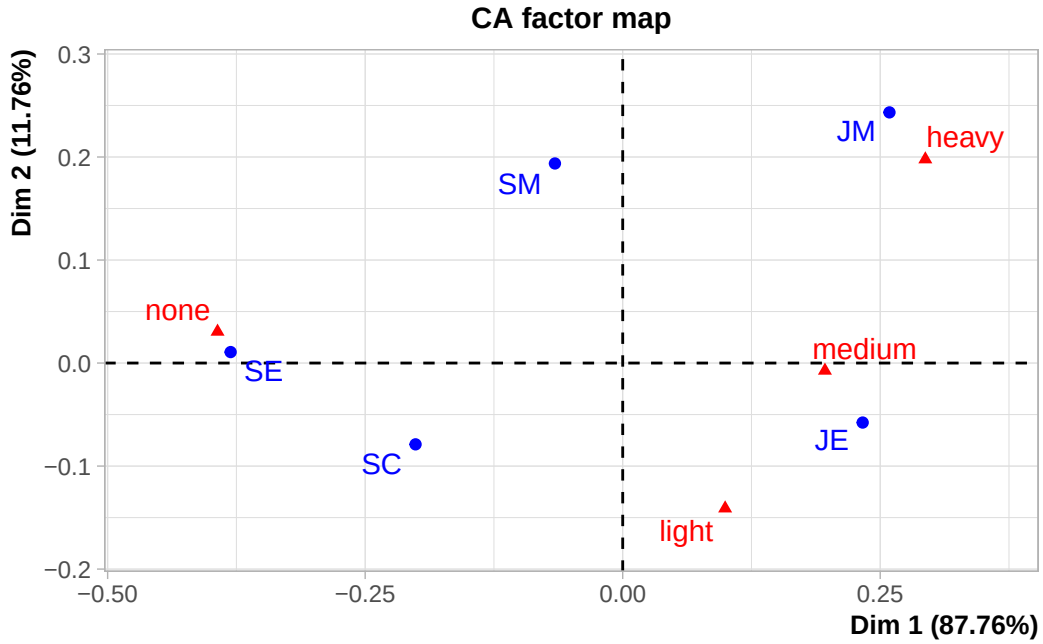
- Les managers junior (`JM`) sont les plus associés au **tabagisme intense** (`heavy`).
- Les employés junior (`JE`) sont plutôt associés au **tabagisme modéré** (`medium`), tandis que les employés senior (`SE`) sont plus associés au **non-fumage** (`none`).

d) Le pourcentage d'inertie expliquée par le premier plan factoriel de l'AFC peut être calcule avec les valeurs propres :

```
Le pourcentage d'inertie expliquée par le premier plan factoriel (Dim 1 + Dim 2) de l'AFC vaut:  99.98
```

- Encore une fois, on peut dire qu'il y a un **effet Guttman** puisque la première dimension explique, à elle seule, explique 99.193% de l'inertie totale.

3. Retrouver ces résultats avec le package FactoMineR et la fonction CA.




```
        eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.0747591059             87.7558731                           87.75587
dim 2 0.0100171805             11.7586535                           99.51453
dim 3 0.0004135741              0.4854734                          100.00000
```

- En réalisant l'analyse factorielle des correspondances avec la fonction `CA` de `FactoMineR`, on obtient les mêmes interprétations qu'avec notre fonction `gsvd`.
- Les légères différences observées dans les valeurs propres proviennent des conventions de normalisation et de pondération des lignes et des colonnes utilisées dans chaque méthode.

Exercice 34. Données `writers`

Il s'agit ici de proposer une méthodologie d'analyse textuelle pour identifier les auteurs de deux fragements de texte anonymes. On connaît pour chacun de ces fragments de texte la fréquence d'apparition de certaines lettres. On supposeégalement que les auteurs de ces textes appartiennent à la liste suivante d'écrivains du 17ème et 18ème siècles : `Charles Darwin`, `René Descartes`, `Thomas Hobbes`, `Mary Shelley` et `Mark Twain`.

Ainsi, 3 échantillons de 1000 caractères de textes de ces auteurs ont été examinés. La fréquence d'apparition de 16 lettres pour chacun de ces 15échantillons est donnée dans un tableau de contingence.

1. Récupérons les données `writers` et chargons les dans R avec la commande `read.csv`. Affichons les données :

Table 4.23 – Jeu de données `writers`

X	B	C	D	F	G	H	I	L	M	N	P	R	S	U	W	Y
CD1	34	37	44	27	19	39	74	44	27	61	12	65	69	22	14	21
CD2	18	33	47	24	14	38	66	41	36	72	15	62	63	31	12	18

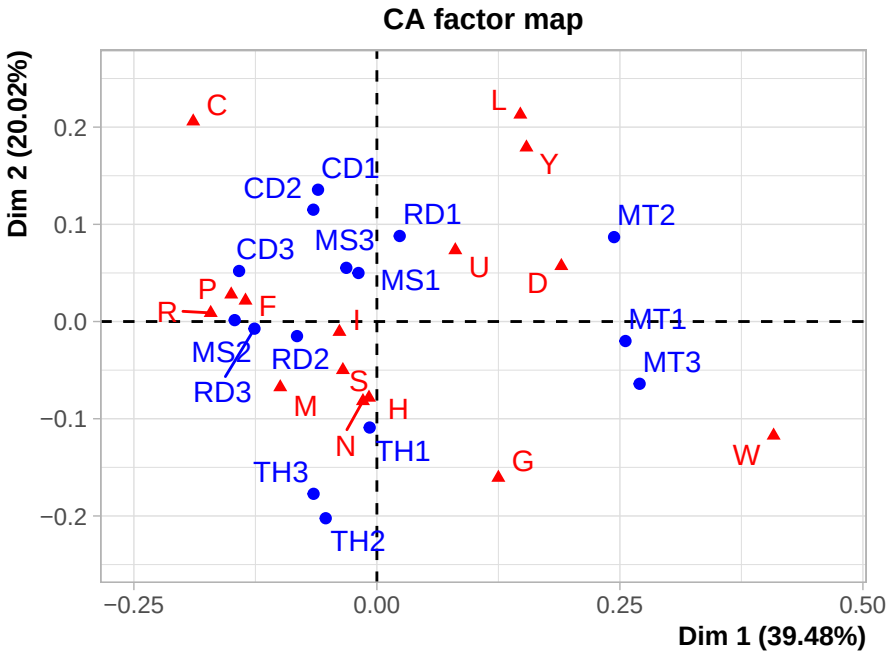
X	B	C	D	F	G	H	I	L	M	N	P	R	S	U	W	Y
CD3	32	43	36	12	21	51	75	33	23	60	24	68	85	18	13	14
RD1	13	31	55	29	15	62	74	43	28	73	8	59	54	32	19	20
RD2	8	28	34	24	17	68	75	34	25	70	16	56	72	31	14	11
RD3	9	34	43	25	18	68	84	25	32	76	14	69	64	27	11	18
TH1	15	20	28	18	19	65	82	34	29	89	11	47	74	18	22	17
TH2	18	14	40	25	21	60	70	15	37	80	15	65	68	21	25	9
TH3	19	18	41	26	29	58	64	18	38	78	15	65	72	20	20	11
MS1	13	29	49	31	16	61	73	36	29	69	13	63	58	18	20	25
MS2	17	34	43	29	14	62	64	26	26	71	26	78	64	21	18	12
MS3	13	22	43	16	11	70	68	46	35	57	30	71	57	19	22	20
MT1	16	18	56	13	27	67	61	43	20	63	14	43	67	34	41	23
MT2	15	21	66	21	19	50	62	50	24	68	14	40	58	31	36	26
MT3	19	17	70	12	28	53	72	39	22	71	11	40	67	20	41	17
TextX1	24	26	80	17	32	91	86	54	32	91	19	58	93	50	58	30
TextX2	19	33	35	22	40	96	116	39	40	129	17	72	104	30	25	24

2. On considère dans un premier temps le tableau de contingence des 15 échantillons dont on connaît les auteurs. Effectuons un test du χ^2 d'indépendence pour répondre à la question : "Les distributions des lettres sont-elles significativement différentes d'un échantillon à l'autre ?"

```
    Pearson's Chi-squared test

data:  tab_echant_lettres
X-squared = 455.18, df = 210, p-value < 2.2e-16
```

- La p-valeur étant inférieure à 0.05, on **rejette** l'hypothèse d'indépendance : Les distributions des lettres **diffèrent significativement** d'un échantillon à l'autre.

3. Puisque les 2 variables sont significativement dépendantes, nous pouvons effectuer une AFC sur les 15 échantillons dont on connaît les auteurs :

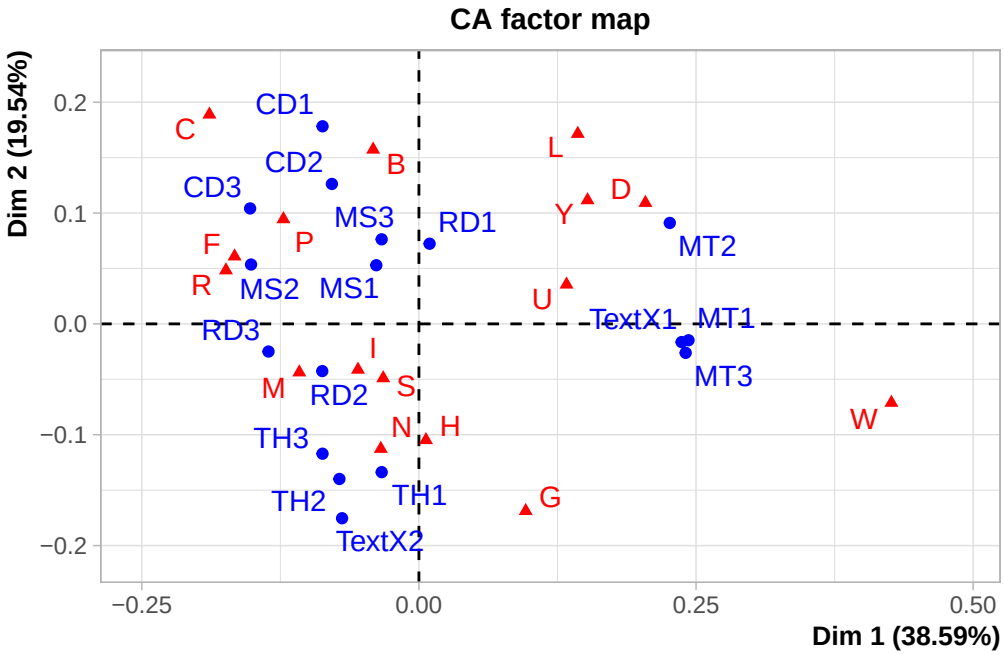


Remarquons que :

- Les lettres **C**, **L**, **Y**, **W** et **G** ne semblent fortement associées à aucun échantillon.
- Les lettres **F** et **R** sont les plus associées au **deuxième échantillon de** `Mary Shelley`.

- Les lettres **I** et **S** sont les plus associées au **deuxième échantillon de** `René Descartes`, tandis que la lettre **U** l'est à son **premier échantillon**.
- Et la lettre **D** est la plus associée au **deuxième échantillon de** `Mark Twain`.

4. Effectuons maintenant une AFC sur les 17 échantillons, en ajoutant les deux textes inconnus en lignes supplémentaires :

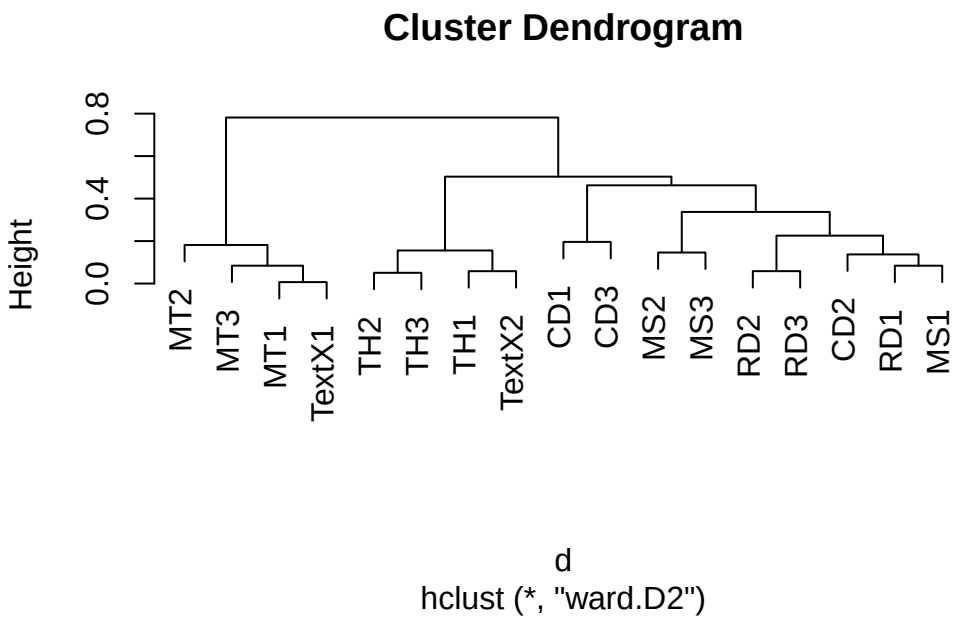


Remarquons que :

- La lettre **B** est la plus associée au **deuxième échantillon de** `Charles Darwin`.
- La lettre **P** est la plus associée au **troisième échantillon de** `Charles Darwin`, tandis que la lettre **C** l'est au **premier échantillon**.
- Les lettres **F** et **R** sont les plus associées au **deuxième échantillon de** `Mary Shelley`.
- Les lettres **M**, **I** et **S** sont particulièrement associées au **deuxième échantillon de** `René Descartes`.
- La lettre **U** est la plus associée au premier texte inconnu.

Par ailleurs, Les échantillons d'un même auteur tendent à se regrouper.

5. Réalisons avec la fonction `hclust` une classification ascendante hiérarchique de Ward des 17 échantillons décrits par leurs coordonnées factorielles sur les 4 premières dimensions de l'AFC.



- La classification ascendante hiérarchique de Ward des 17 échantillons confirme encore une fois que les échantillons d'un même auteur tendent à se regrouper :
- `TH2` est plus proche de `TH3` que de `TH1`.
- `RD2` est plus proche de `RD3` que de `RD1`.
- Si on coupe l'arbre pour avoir 4 branches, tous les échantillons d'un même auteur se retrouvent presque tous dans un même groupe.
- Remarquons aussi que le premier texte `TextX1` d'auteur inconnu est plus proche de l'échantillon `MT1` de `Mark Twain`, tandis que le second est plus proche de l'échantillon `TH1` de `Thomas Hobbes`.

Affichons la partition en 4 classes :

```
Clusters:  1 2 1 2 2 2 3 3 3 2 2 2 4 4 4 4 3
```

5 Travail personnel 5 (ACM)

Exercice 27. Données `chiens`

1. Récupérons les jeux de données `chiens`. Il s'agit de données fictives ou 27 races de chiens sont décrites avec 7 variables qualitatives. Affichons la classe de cet objet et les premières données :

```
Classe de l'objet 'chiens':  data.frame
```

Table 5.1 – Aperçu des données `chiens`

	taille	poids	velocite	intellig	affect	agress	fonction
beauceron	T++	P+	V++	I+	Af+	Ag+	Utilite

	taille	poids	velocite	intellig	affect	agress	fonction
basset	T-	P-	V-	I-	Af-	Ag+	Chasse
ber_allem	T++	P+	V++	I++	Af+	Ag+	Utilite
boxer	T+	P+	V+	I+	Af+	Ag+	Compagnie
bull-dog	T-	P-	V-	I+	Af+	Ag-	Compagnie
bull-mass	T++	P++	V-	I++	Af-	Ag+	Utilite

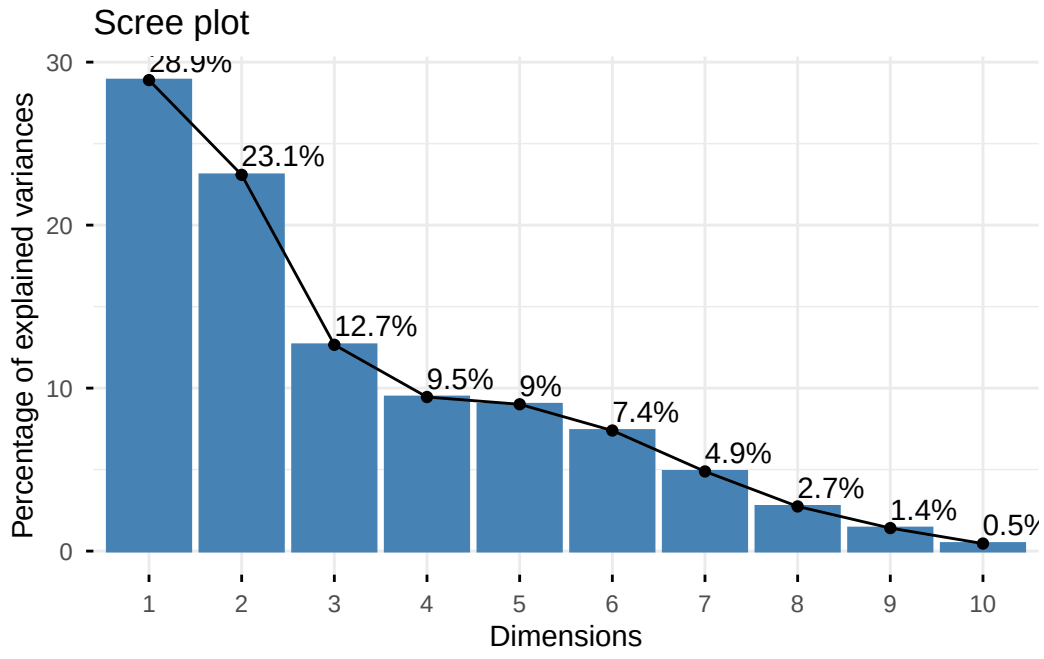
2. Créons une matrice H contenant la description des n = 27 races canines sur uniquement les p = 6 premières variables :

Table 5.2 – Aperçu des données `chiens` avec 6 premières variables

	taille	poids	velocite	intellig	affect	agress
beauceron	T++	P+	V++	I+	Af+	Ag+
basset	T-	P-	V-	I-	Af-	Ag+
ber_allem	T++	P+	V++	I++	Af+	Ag+
boxer	T+	P+	V+	I+	Af+	Ag+
bull-dog	T-	P-	V-	I+	Af+	Ag-
bull-mass	T++	P++	V-	I++	Af-	Ag+

3. On va effectuer l'ACM de cette matrice H.

d) Représentons dans un diagramme en barre les pourcentages d'inertie expliquée par les dimensions de l'ACM.



- Les axes à retenir seraient les 3 premières : elles expliquent 64.7% de l'inertie totale.

e) Déterminons les matrices X et Y des coordonnées factorielles des races de chiens et des modalités des variables qualitatives sur les $k = 3$ premières dimensions. Modifions les noms des lignes et des colonnes dans X et Y afin qu'ils soient parlants.

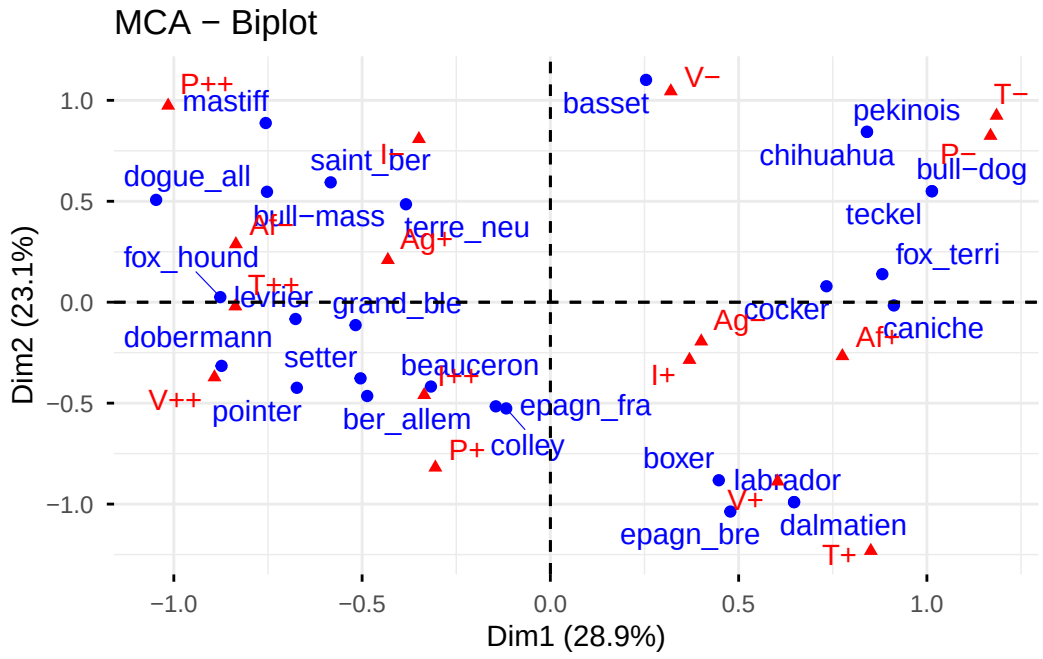
Table 5.3 – Aperçu des Coordonnées des individus (races) : Données H

	Dim1	Dim2	Dim3
beauceron	-0.3172001	-0.4177013	-0.1014677
basset	0.2541098	1.1012270	-0.1907010
ber_allem	-0.4863955	-0.4644496	-0.4981339
boxer	0.4473649	-0.8817779	0.6920158
bull-dog	1.0133522	0.5498795	-0.1634232
bull-mass	-0.7525745	0.5469118	0.4975731

Table 5.4 – Aperçu des Coordonnées des modalités : Données H

	Dim1	Dim2	Dim3
T-	1.1849557	0.9238965	-0.6159996
T+	0.8510880	-1.2317197	1.0160518
T++	-0.8366753	-0.0205785	-0.0512174
P-	1.1689180	0.8243446	-0.3587704
P+	-0.3054053	-0.8188757	-0.2312721
P++	-1.0151341	0.9739006	1.2215945

f) Faisons un biplot des individus et des modalités dans le premier plan factoriel (1,2) :



Remarquons, par exemple, que :

- L'axe 1 oppose les races de petite taille et de petit poids aux races de grande taille et de poids élevé.
- L'axe 1 oppose aussi les races les moins agressives aux races les plus agressives.
- Quant à l'axe 2, il oppose les races les plus intelligentes (exemple : beauceron) aux races les moins intelligentes (exemple : saint_ber).
- Il oppose aussi les races les moins affectueuses (bull_mass) aux races les plus affectueuses (caniche) ; les moins rapides (basset) aux plus rapides (dobermann, dalmatien).
- Par ailleurs, les races comme `boxer`, `labrador`, `dalmatien` et `epagn_bre` ont des profils similaires.
- Les races comme `bulldog`, `teckel`, `chihuahua` et `pekinois` ont des profils similaires.

g) Utiliser la relation quasi-barycentrique pour retrouver les coordonnées factorielles de la modalité T++ à partir des coordonnées factorielles des races de chiens.

Table 5.5 – Coordonnées factorielles de la modalité `T++` (barycentre)

	x
Dim1	-0.5806347
Dim2	-0.0127642
Dim3	-0.0235240

h) Affichons les rapports de corrélation entre la variable taille avec la première composante principale et entre la variable taille et la seconde composante principale :

Table 5.6 – Rapport de corrélation entre `taille` et les 2 premières dimensions

	x
Dim 1	0.8870733
Dim 2	0.5024857

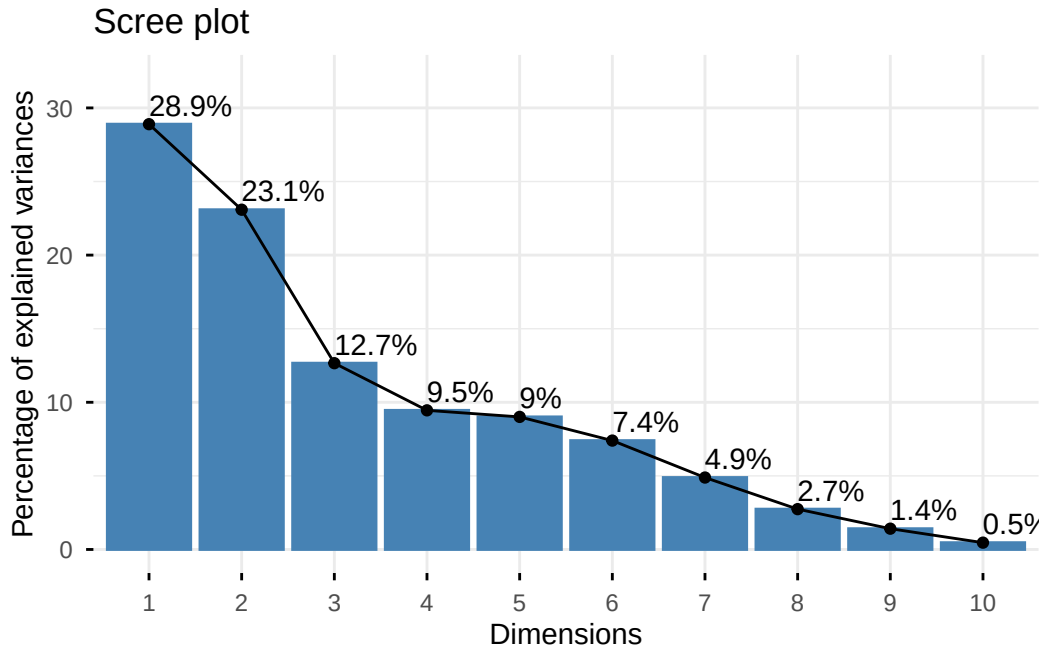
- La variable `taille` est donc fortement corrélée avec la première composante principale mais faiblement corrélée avec la seconde composante principale.

4. On va maintenant réaliser l'Analyse des Correspondances Multiples (ACM) des données sur les races canines en mettant la variable fonction en illustratif.

a) Faisons l'ACM en indiquant la variable `fonction` comme `quali.sup`.

b) Retrouvons les résulats numériques et les graphiques de la question 2.

1. Représentons dans un diagramme en barre les pourcentages d'inertie expliquée par les dimensions de l'ACM.



- Les axes à retenir seraient les 3 premières : elles expliquent 64.7% de l'inertie totale.

2. Déterminons les matrices X et Y des coordonnées factorielles des races de chiens et des modalités des variables qualitatives sur les $k = 3$ premières dimensions. Modifions les noms des lignes et des colonnes dans X et Y afin qu'ils soient parlants.

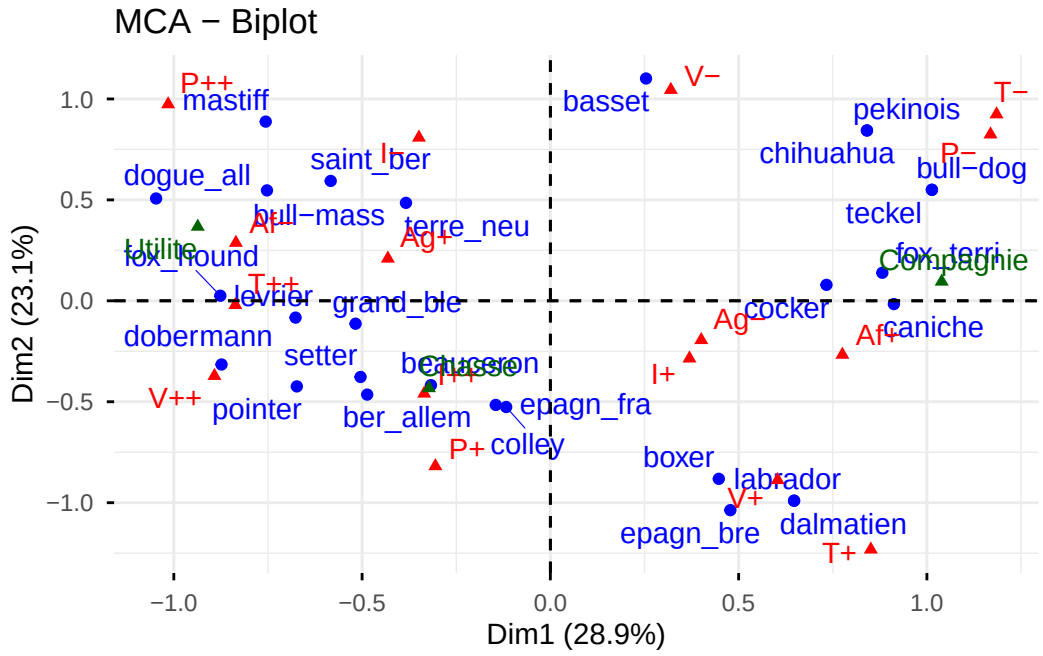
Table 5.7 – Aperçu des Coordonnées des individus (races) sur les 3 premières dimensions

	Dim1	Dim2	Dim3
beauceron	-0.3172001	-0.4177013	-0.1014677
basset	0.2541098	1.1012270	-0.1907010
ber_allem	-0.4863955	-0.4644496	-0.4981339
boxer	0.4473649	-0.8817779	0.6920158
bull-dog	1.0133522	0.5498795	-0.1634232
bull-mass	-0.7525745	0.5469118	0.4975731

Table 5.8 – Aperçu des Coordonnées des modalités sur les 3 premières dimensions

	Dim1	Dim2	Dim3
T-	1.1849557	0.9238965	-0.6159996
T+	0.8510880	-1.2317197	1.0160518
T++	-0.8366753	-0.0205785	-0.0512174
P-	1.1689180	0.8243446	-0.3587704
P+	-0.3054053	-0.8188757	-0.2312721
P++	-1.0151341	0.9739006	1.2215945

3. Faisons un biplot des individus et des modalités dans le premier plan factoriel (1,2) :



Remarquons, par exemple, que :

- L'axe 1 oppose les races de petite taille et de petit poids, très souvent de compagnie, aux races de grande taille et de poids élevé, souvent de chasse ou d'utilité.
- L'axe 1 oppose aussi les races les moins agressives (de compagnie) aux races les plus agressives (de chasse).
- Quant à l'axe 2, il oppose les races les plus intelligentes (de chasse) aux races les moins intelligentes (d'utilité).
- Par ailleurs, les races comme `boxer`, `labrador`, `dalmatien` et `épagneul epagn_bre` ne sont fortement associées à aucune fonction.
- Alors que les races comme `bulldog`, `teckel`, `chihuahua` et `pekinois` sont fortement associées à la fonction d'animaux de compagnie.
- Les races comme les `bergers allemands`, `épagneul français`, `colley` et `beauceron` sont fortement associées à la fonction d'animaux de chasse, tandis que les `fox_hound`, `bull_mass` et `dogue_all` sont plutôt associées a la fonction d'utilité.

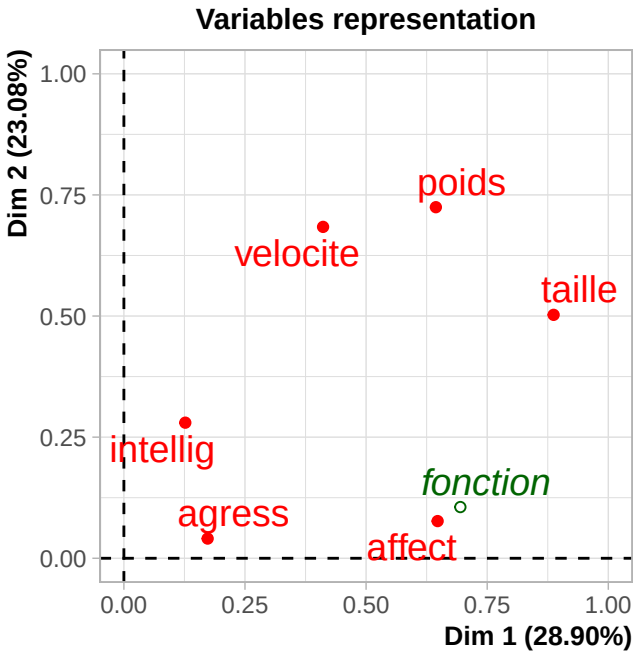
c) Retrouvons les rapports de corrélations entre les variables qualitatives et les deux premières composantes principales :

Table 5.9 – Rapports de corrélations entre les variables qual. et les 2 premières CP

	Dim 1	Dim 2
taille	0.8870733	0.5024857
poids	0.6440465	0.7246877
velocite	0.4111741	0.6840074
intellig	0.1267635	0.2798701
affect	0.6476559	0.0767360
agress	0.1729238	0.0406369

- Les variables `taille`, `affect` sont plus fortement corrélées avec la première composante principale qu'avec la seconde.
- Les variables `poids` et `velocite` sont plus fortement corrélées avec la seconde composante principale qu'avec la première.

Faisons le plot des variables en fonction de ces rapports de corrélation en utilisant la fonction `plot.MCA`.

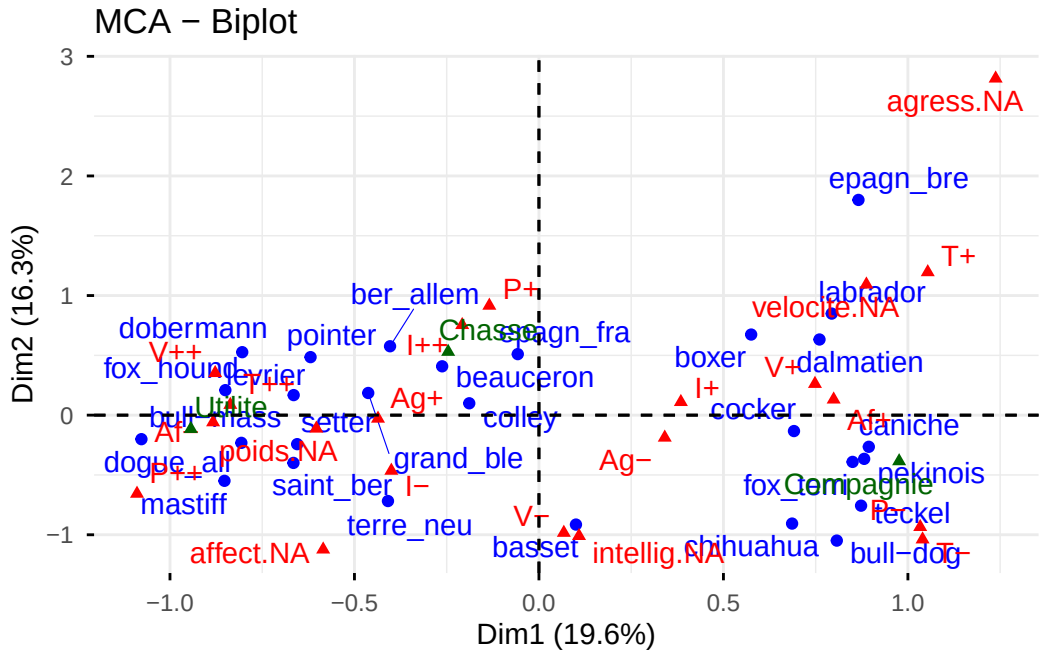


d) Ajoutons des données manquantes dans les données et appelons ce nouveau jeu `chiensNA` :

Table 5.10 – Aperçu des données `chiensNA`

	taille	poids	velocite	intellig	affect	agress	fonction
beauceron	T++	P+	V++	I+	Af+	Ag+	Utilite
basset	T-	P-	V-	I-	Af-	Ag+	Chasse
ber_allem	T++	P+	V++	I++	Af+	Ag+	Utilite
boxer	T+	P+	V+	I+	Af+	Ag+	Compagnie
bull-dog	T-	P-	V-	NA	Af+	Ag-	Compagnie
bull-mass	T++	P++	V-	I++	Af-	Ag+	Utilite

e) Faisons l'ACM du jeu de données `chiensNA` :



- Dans la fonction `MCA` de `FactoMineR`, les valeurs manquantes sont traitées comme des modalités spécifiques (`nom_variable.NA`), ce qui permet de conserver tous les individus dans l'analyse et d'évaluer l'impact des données manquantes sur la structure factorielle.
- Par ailleurs, ces valeurs manquantes sont ignorées dans le calcul des coordonnées des individus : l'ACM est réalisée avec les informations disponibles, sans supprimer entièrement les lignes contenant des NAs.

5. On va maintenant comparer l'ACM et l'AFC dans le cas particulier de deux variables qualitatives.

a) Avec la fonction `CA` de `FactoMineR`, effectuons l'AFC du tableau de contingence croisant les variables `taille` et `poids`.

b) Avec la fonction `MCA`, effectuons l'ACM des deux premières colonnes des données `chiens` :

c) Comparons les valeurs propres des deux analyses et vérifions que nous retrouvons les relations du cours :

Table 5.11 – Valeurs propres AC

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.8606329	91.742589	91.74259
dim 2	0.0774624	8.257411	100.00000

```
Somme des valeurs propres AC : 0.9380952

chi²/n : 0.9380952

Nombre de valeurs propres AC =  2

Nombre d'axes AC = min(r,c) - 1 =  2
```

Table 5.12 – Valeurs propres ACM

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.9638515	48.192575	48.19258
dim 2	0.6391603	31.958016	80.15059
dim 3	0.3608397	18.041984	98.19258
dim 4	0.0361485	1.807425	100.00000

```
Somme des valeurs propres ACM : 2

(m1 + m2 - p)/p :  2

Nombre de valeurs propres ACM =  4
```

Toutes les relations du cours ont été retrouvées :

- La somme des valeurs propres de l'AFC est égale à $\frac{\chi^2}{n}$.
- Le nombre d'axes de l'AFC est égal à $\min(r, c) - 1$, qui est aussi le nombre de valeurs propres.
- La somme des valeurs propres de l'AFCM est égale à $\frac{(m_1+m_2+\cdots+m_p)-p}{p}$
- Le nombre d'axes de l'AFCM est égal au double du nombre d'axes de l'AFC.

Exercice 28. ACM avec données manquantes et choix du nombre de composantes

Le package `missMDA` offre des outils puissants pour gérer les **données manquantes** dans les analyses multivariées, notamment en **ACP** (Analyse en Composantes Principales) et **ACM** (Analyse des Correspondances Multiples).

Il permet :

- d'estimer automatiquement les valeurs manquantes,
- de choisir le nombre de composantes par **validation croisée**,
- et d'intégrer ce traitement dans une démarche d'analyse complète.

1. Estimation des données manquantes avec `(estim_ncpMCA / estim_ncpPCA)`

Ces fonctions utilisent une procédure de **validation croisée** pour déterminer le nombre optimal de composantes à retenir, afin de reconstruire les valeurs manquantes de manière cohérente avec la structure des données.

Pour illustrer leur utilisation en ACM, nous allons l'appliquer à notre jeu de données `chiensNA`.

Table 5.13 – Aperçu des données `chiensNA` avant imputation

	taille	poids	velocite	intellig	affect	agress	fonction
beauceron	T++	P+	V++	I+	Af+	Ag+	Utilite
basset	T-	P-	V-	I-	Af-	Ag+	Chasse
ber_allem	T++	P+	V++	I++	Af+	Ag+	Utilite
boxer	T+	P+	V+	I+	Af+	Ag+	Compagnie
bull-dog	T-	P-	V-	NA	Af+	Ag-	Compagnie
bull-mass	T++	P++	V-	I++	Af-	Ag+	Utilite

```
Nombre optimal de dimensions pour la reconstruction des NAs:  3
```

La fonction renvoie le nombre optimal **3** de dimensions `ncp` à utiliser pour l'imputation.

2. Imputation des valeurs manquantes (`imputeMCA, imputePCA`)

Les valeurs manquantes sont estimées en projetant les individus dans l'espace factoriel défini par les composantes principales de l'ACP ou de l'ACM. Cette projection utilise les relations entre variables pour calculer des valeurs cohérentes avec la structure globale des données. Ainsi, le jeu de données imputé conserve les dépendances entre variables tout en permettant de réaliser l'analyse factorielle complète.

Dans notre exemple, on obtient le jeu de données complétées suivant :

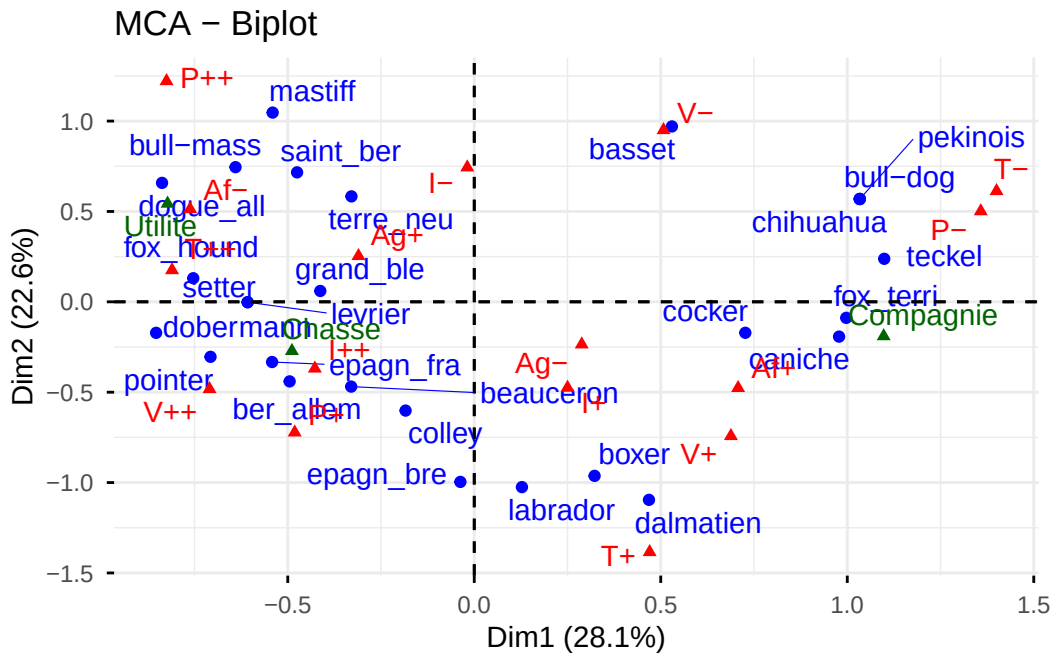
Table 5.14 – Aperçu des données `chiensNA` après imputation

	taille	poids	velocite	intellig	affect	agress	fonction
beauceron	T++	P+	V++	I+	Af+	Ag+	Utilite
basset	T-	P-	V-	I-	Af-	Ag+	Chasse
ber_allem	T++	P+	V++	I++	Af+	Ag+	Utilite
boxer	T+	P+	V+	I+	Af+	Ag+	Compagnie
bull-dog	T-	P-	V-	I-	Af+	Ag-	Compagnie
bull-mass	T++	P++	V-	I++	Af-	Ag+	Utilite

Les NAs sont remplacés par des valeurs estimées, prêtes à être utilisées pour une ACM complète.

3. Réalisation d'ACM sur les données complétées

A partir de là, on peut réaliser une Analyse des Correspondances Multiples sur notre jeu de données complétées en utilisant la fonction `MCA` de `FactoMineR` :



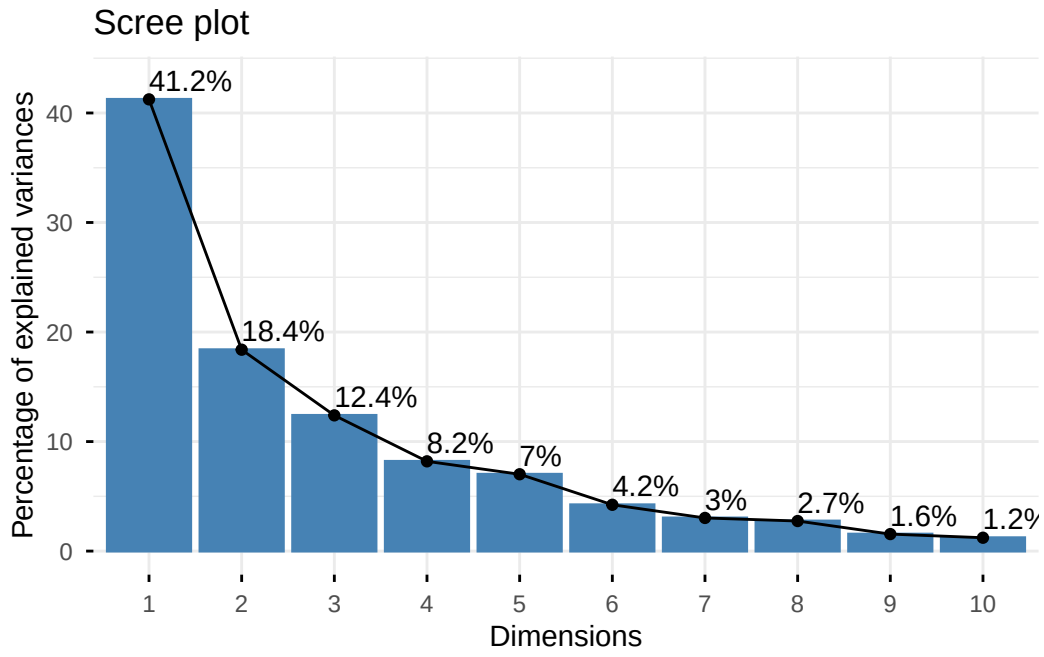
- Comme François Husson l'a précisé, il faut noter que si on a une faible proportion de données manquantes, les pourcentages d'inertie des composantes principales peuvent être légèrement surestimés ou sous-estimés. Et si la proportion de données manquantes est importante, alors les pourcentages seront fortement surestimés ou sous-estimés.

- Si on compare l'ACM sur les données originales complètes `chiens` et l'ACM sur les données imputées `chiensNA.complet`, les pourcentages d'inertie des 2 premières composantes principales C1 et C2 passent respectivement de 28.9% à 28.1% et de 23.1% à 22.8%. (donc légèrement sous-estimés)

6 Travail personnel 6 (Classification)

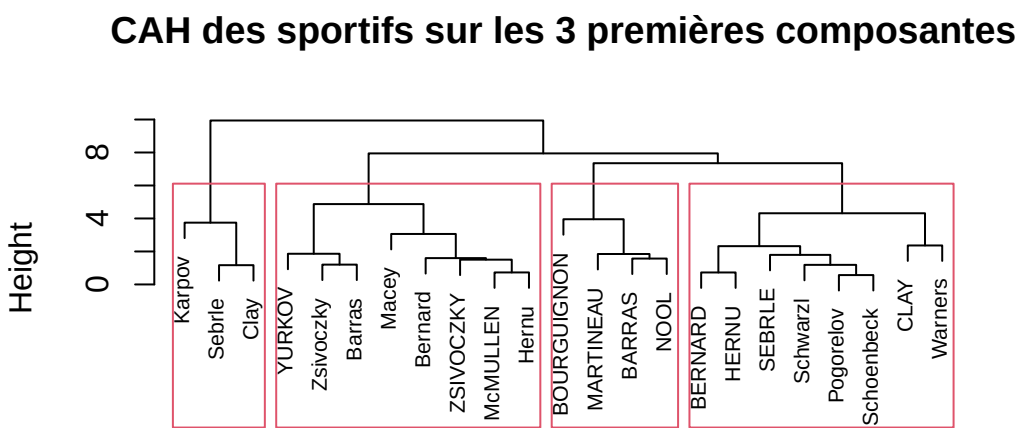
Exercice 29.1 CAH sur les données `decathlon2`

On va d'abord faire une ACP pour réduire le mnombre de dimensions et ensuite, on va classifier sur les dimensions retenues.

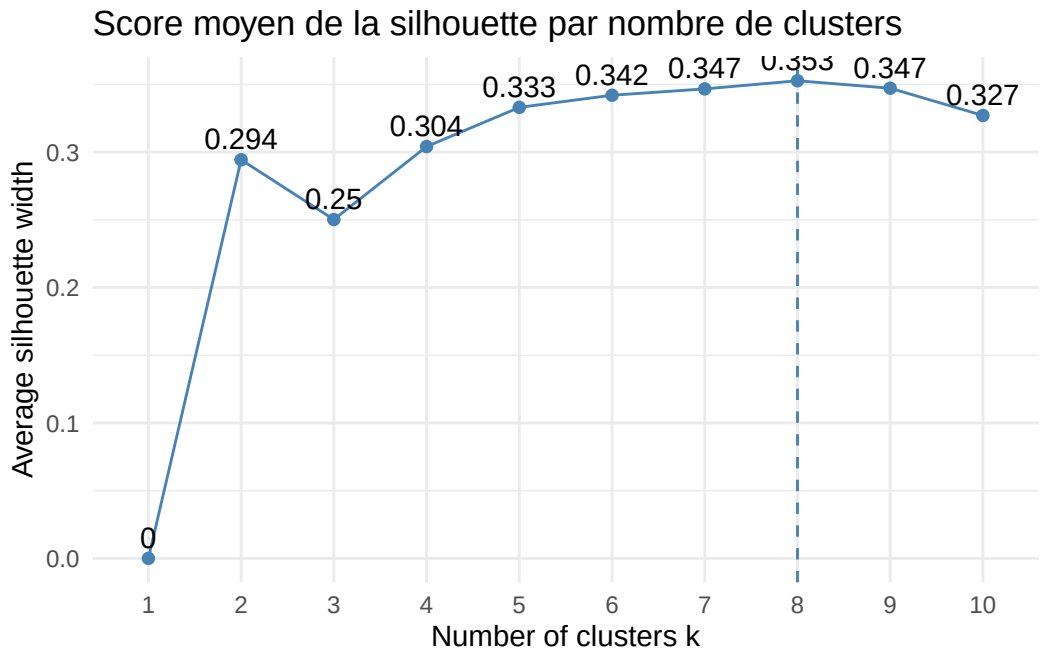


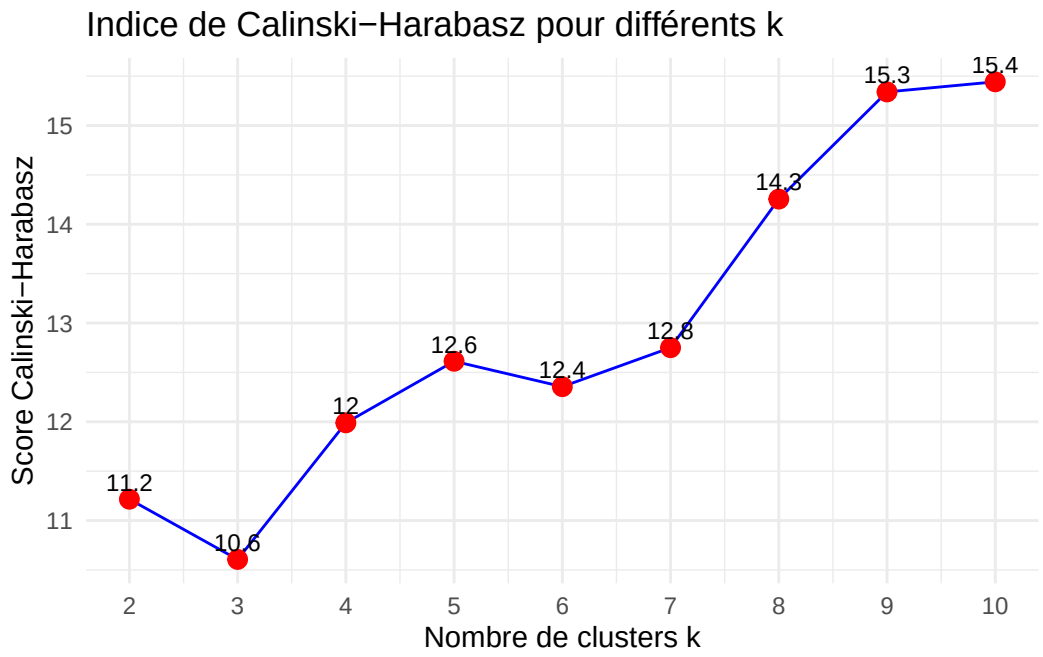
- Les 3 premières composantes principales expliquent 72.01885% de la variance totale. C'est un pourcentage acceptable.

Réalisons une CAH sur les données en ne retenant que les 3 premières composantes principales :



Faisons la consolidation par k-means :

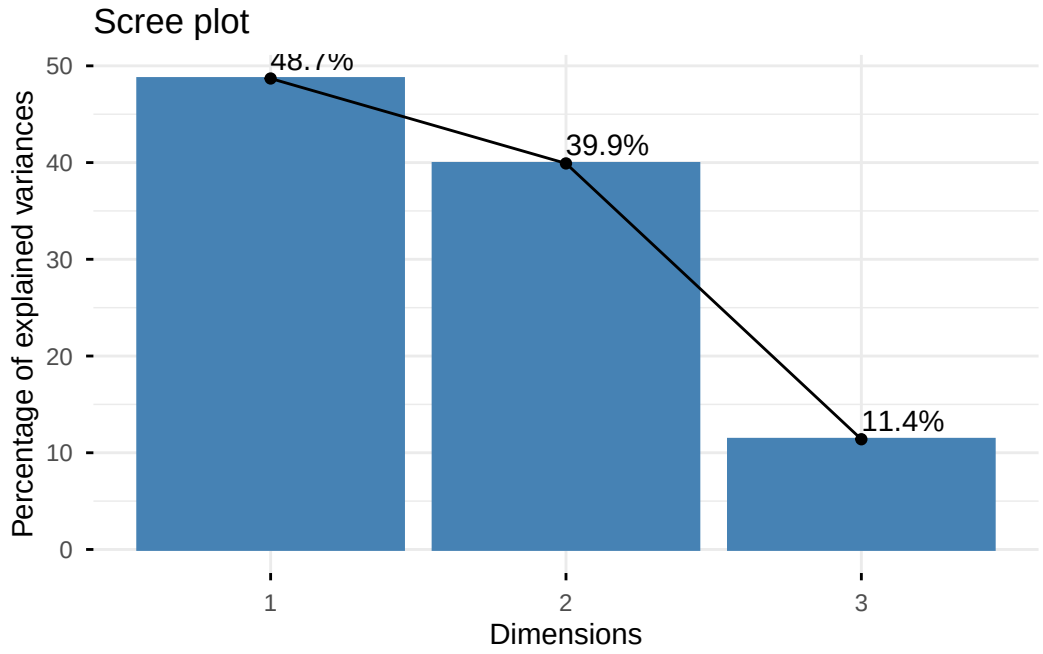




- La première méthode, basée sur la méthode de la silhouette, suggère un nombre optimal $k = 8$ clusters mais le score 0.35 reste très faible.
- En plus, avoir 8 clusters impliquerait d'avoir plusieurs classes avec un seul individu : ce qui compliquerait interprétabilité des classes.
- Par ailleurs, l'indice de Calinski-Harabasz ne suggère rien car il n'y a pas de pic.

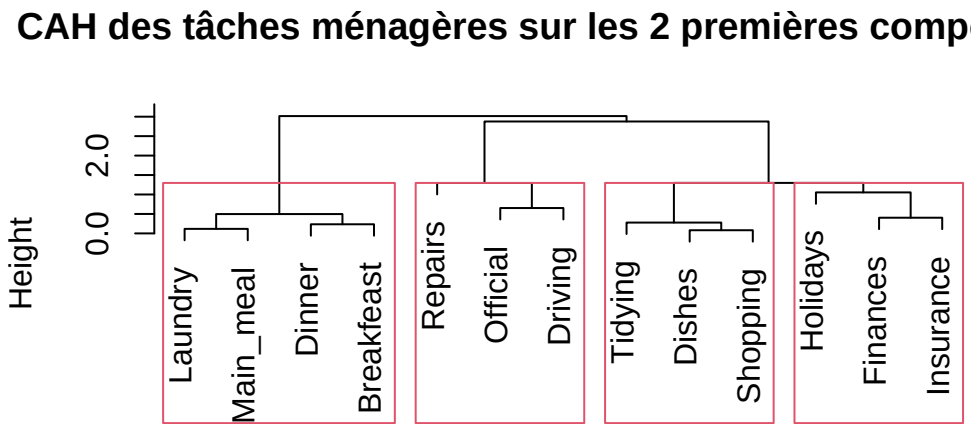
Exercice 29.2 CAH sur les données `housetasks`

On va d'abord faire une AFC pour réduire le nombre de dimensions et ensuite, on va classifier sur les dimensions retenues.

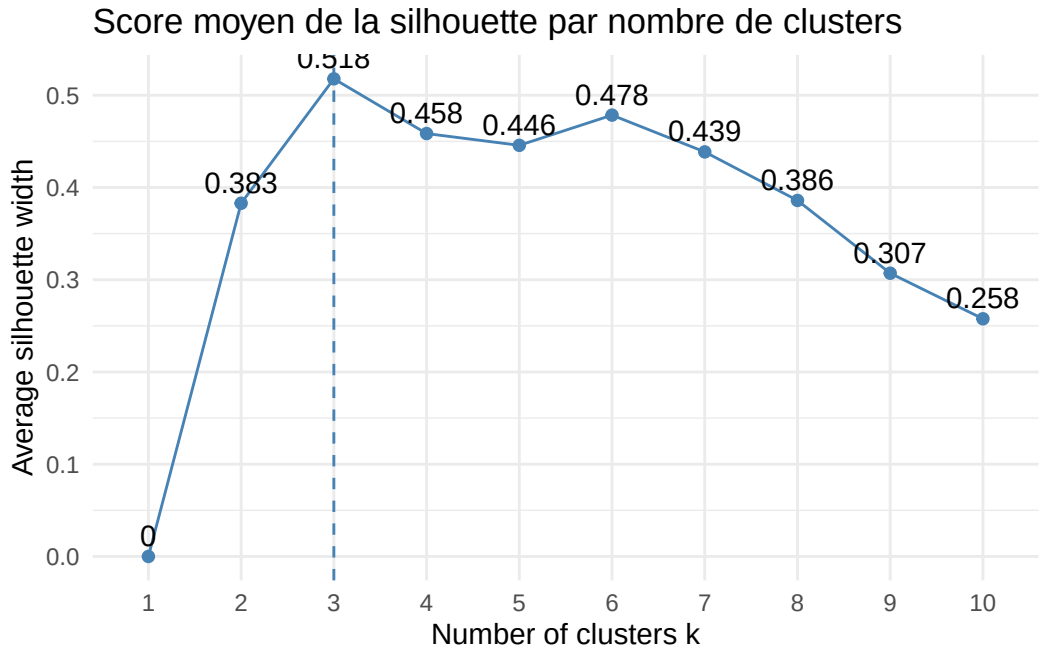


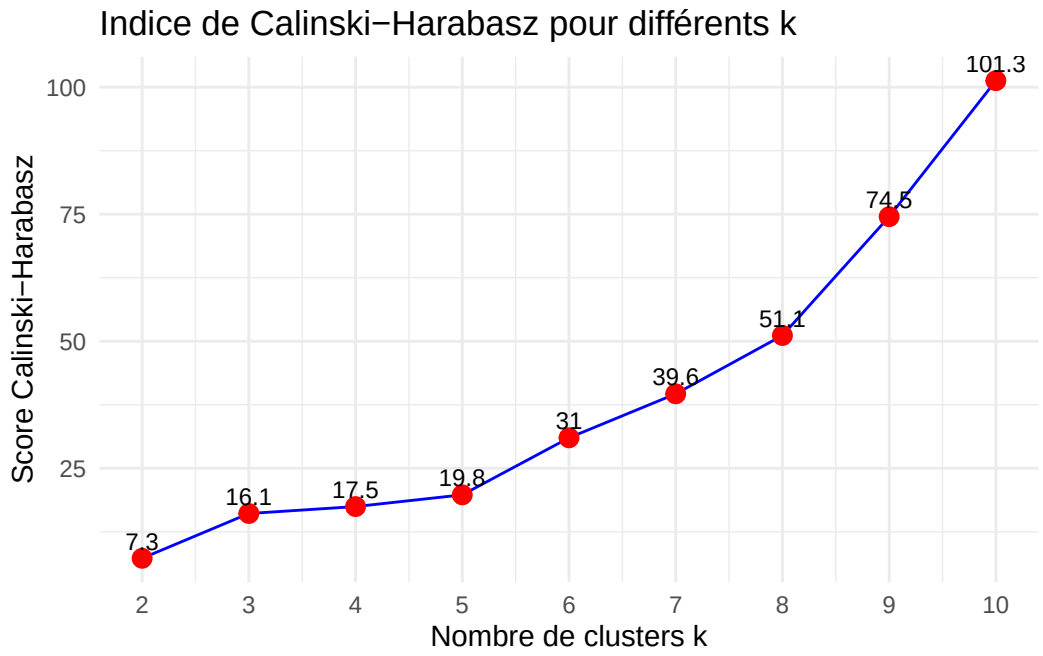
- Ici, on retient les 2 premières dimensions. Retenir les 3 voudrait dire qu'on n'a fait aucune réduction.

Réalisons une CAH sur les données en ne retenant que les 2 premières composantes principales :



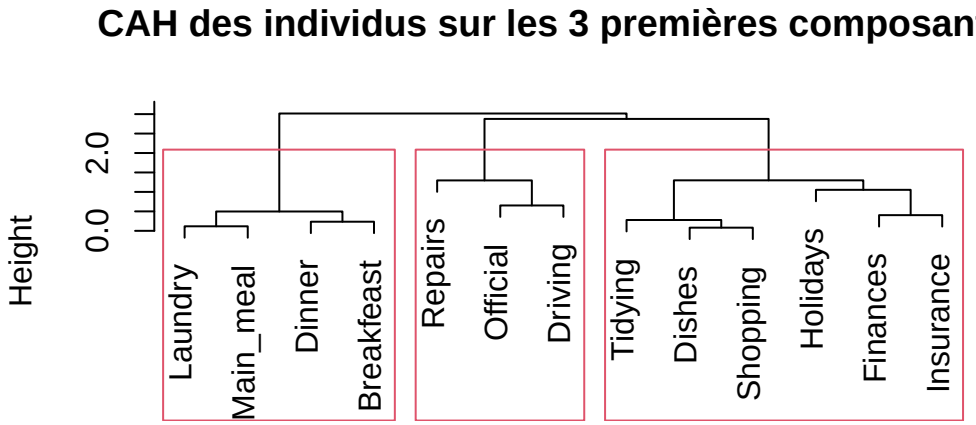
Faisons la consolidation par k-means :





- La première méthode, basée sur la méthode de la silhouette, suggère un nombre optimal $k = 3$ clusters. Le score 0.518 n'est pas fort mais on pourrait quand même essayer.
- L'indice de Calinski-Harabasz ne montre pas de pic et donc, on essaiera avec la proposition de la première méthode.

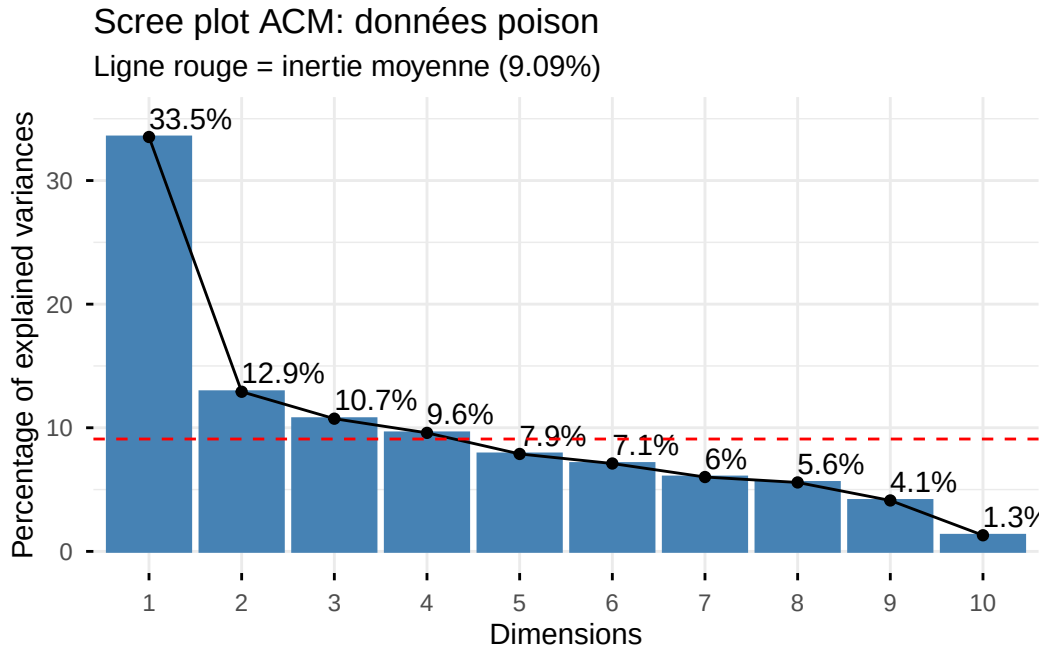
Faisons dpnc une classification avec $k = 3$ clusters :



Exercice 29.3 CAH sur les données `poison`

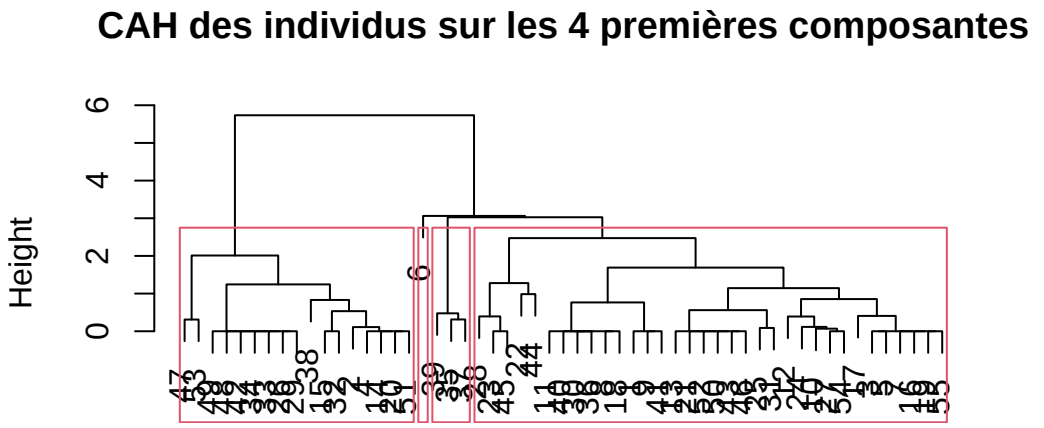
On va d'abord faire une AFCM pour réduire le nombre de dimensions et ensuite, on va classifier sur les dimensions retenues.

Age	Time	Sick	Sex	Nausea	Vomiting	Abdominal	Fever	Diarrhae	Potato	Fish	Mayo	Courgette	Cheese	Icecream
9	22	Sick_y	F	Nausea_y	Vomit_y	Abdo_y	Fever_y	Diarrhea_y	Potato_y	Fish_y	Mayo_y	Courg_y	Cheese_y	Icecream_y
5	0	Sick_n	F	Nausea_n	Vomit_n	Abdo_n	Fever_n	Diarrhea_n	Potato_y	Fish_y	Mayo_y	Courg_y	Cheese_n	Icecream_y
6	16	Sick_y	F	Nausea_n	Vomit_y	Abdo_y	Fever_y	Diarrhea_y	Potato_y	Fish_y	Mayo_y	Courg_y	Cheese_y	Icecream_y
9	0	Sick_n	F	Nausea_n	Vomit_n	Abdo_n	Fever_n	Diarrhea_n	Potato_y	Fish_y	Mayo_n	Courg_y	Cheese_y	Icecream_y
7	14	Sick_y	M	Nausea_n	Vomit_y	Abdo_y	Fever_y	Diarrhea_y	Potato_y	Fish_y	Mayo_y	Courg_y	Cheese_y	Icecream_y
72	9	Sick_y	M	Nausea_n	Vomit_n	Abdo_y	Fever_y	Diarrhea_y	Potato_y	Fish_n	Mayo_y	Courg_y	Cheese_y	Icecream_y

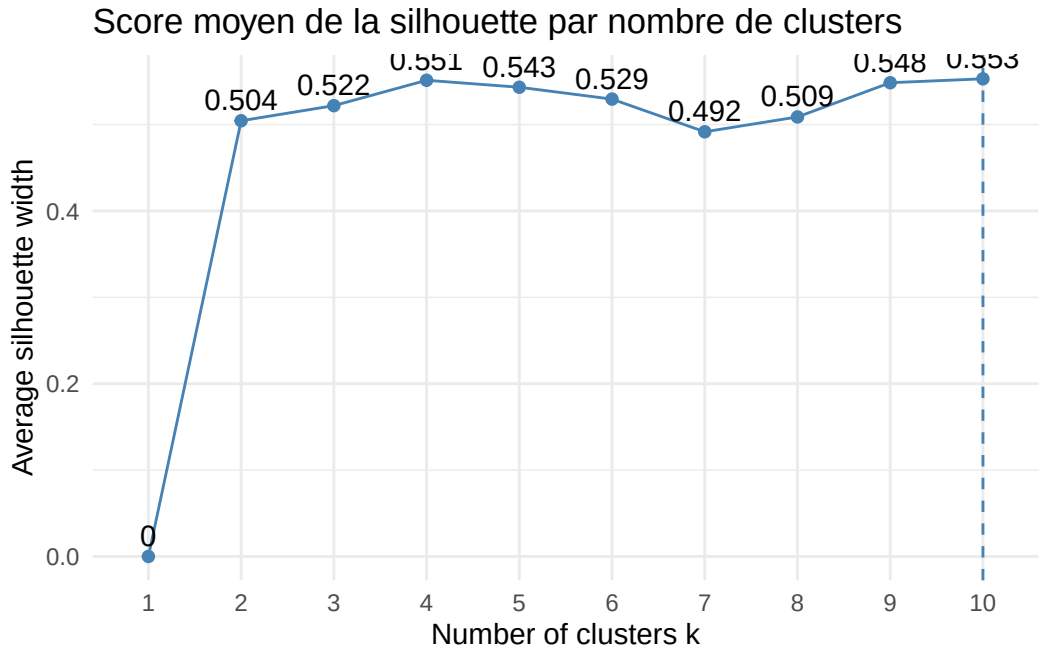


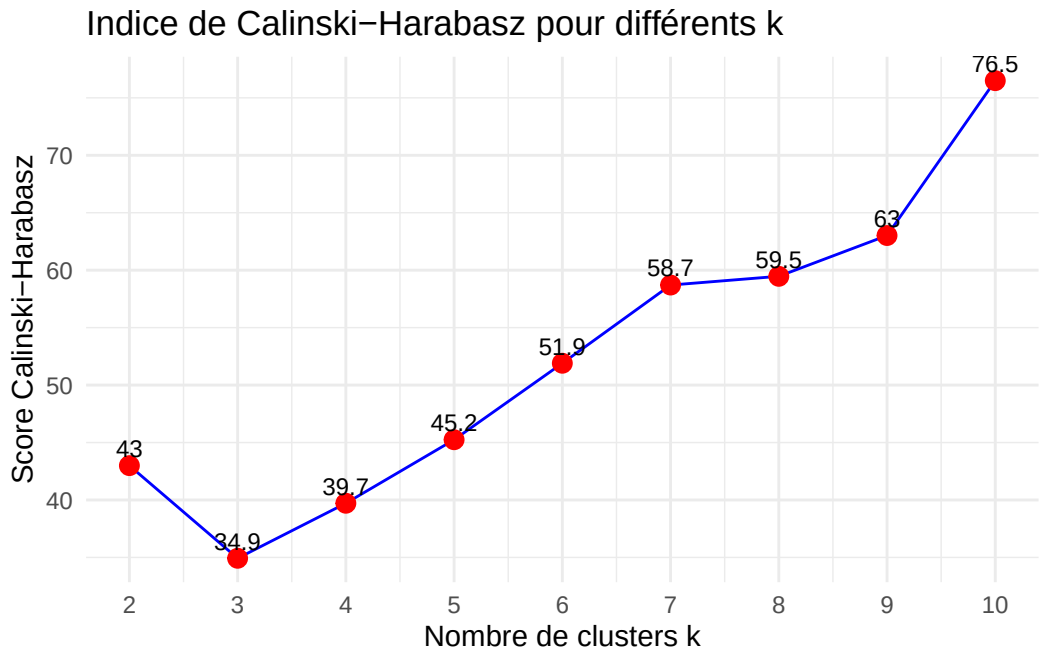
- Le nombre d'axes retenus est déterminé à partir de l'inertie cumulée et de la comparaison à l'inertie moyenne. Les quatre premières dimensions, expliquant environ 67 % de l'inertie totale, sont conservées pour l'analyse.

Réalisons une CAH sur les données en ne retenant que les 4 premières composantes principales :



Faisons la consolidation par k-means :





- **Aucune des deux méthodes** ne suggère un nombre optimal de clusters car il n'y a pas de pic.

Exercice 29.3 CAH sur les données `OCDE`

On va d'abord faire une AFDM pour réduire le nombre de dimensions et ensuite, on va classifier sur les dimensions retenues.

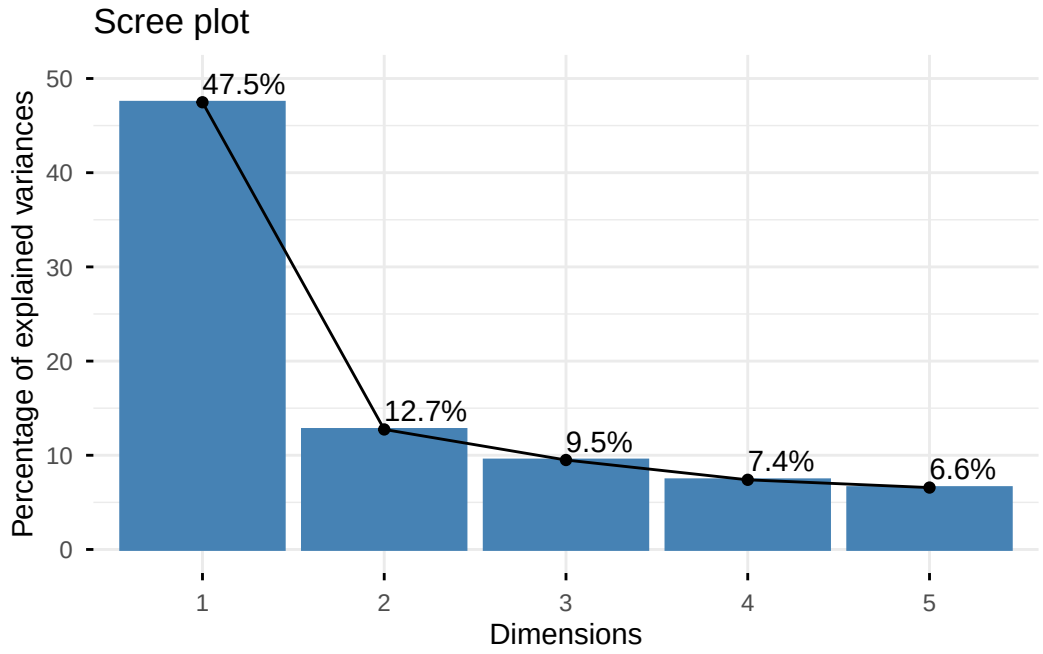
Table 6.2 – Structure des données

Variable	Type	Valeurs
NoPays	int	1 2 3 4 5 6 7 8 9 10 ...
Pays	chr	: chr "Australie" "Autriche" "Belgique" "Canada" ...
popul	num	: num 19.75 8.05 10.33 31.41 10.21 ...
CO2	num	: num 342.9 66.1 112.5 531.9 115 ...
TPES	num	: num 112.7 30.4 56.9 250 41.7 ...
electr	num	187.2 63.8 87.3 555 60.4 ...
import	num	0 19 15 30.5 10.1 7 11.9 6.2 45.4 4.2 ...
export	num	0 13.4 8.2 36.4 26.3 15.6 7 72.2 54.1 2.1 ...
fuel	num	: num 169.2 22.4 34.3 158 50.5 ...
nuclear	num	0 0 44.8 70.3 24.4 ...
hydro	num	: num 18 31.9 1.3 332.6 1.8 ...
geoth	num	: num 0 3.9 0.1 0 0 5.6 0.1 0 0 1.1 ...
zone	chr	: chr "pacasi" "europ" "europ" "ameri" ...
clim	chr	: chr "aride" "temp" "temp" "arct" ...

Table 6.3 – Composantes principales de l'AFDM des données `OCDE`

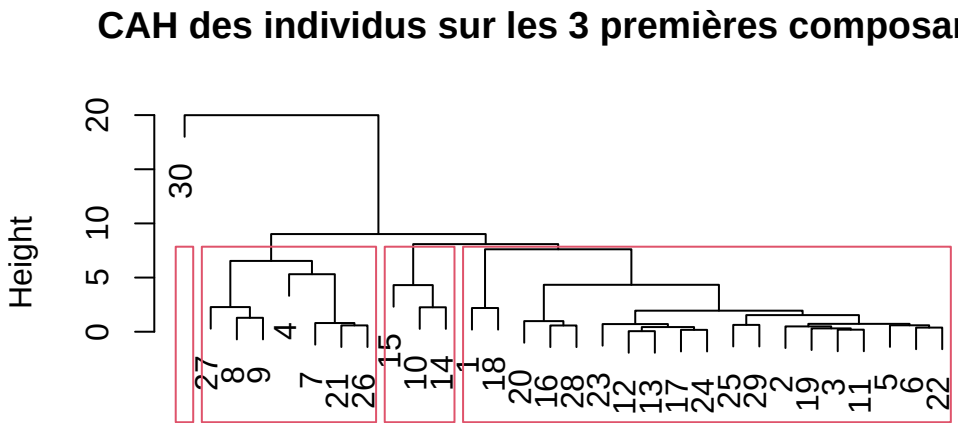
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	7.596267	47.476671	47.47667
comp 2	2.038742	12.742136	60.21881
comp 3	1.519618	9.497616	69.71642
comp 4	1.183446	7.396536	77.11296
comp 5	1.051497	6.571856	83.68481

4. Visualisons les composantes principales et leur pourcentage de variance expliquée :

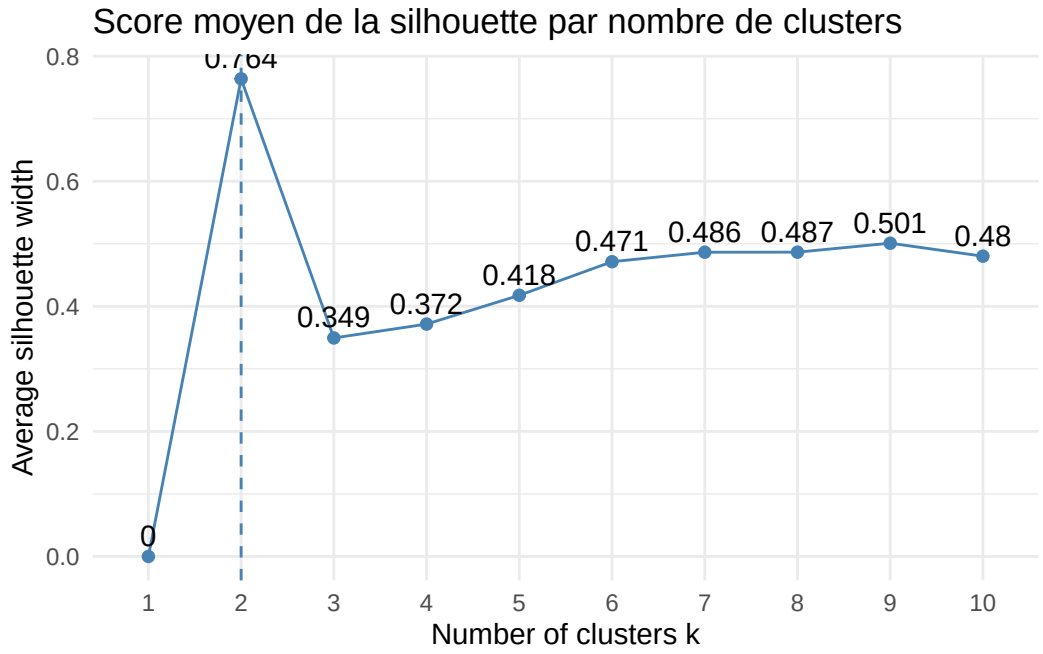


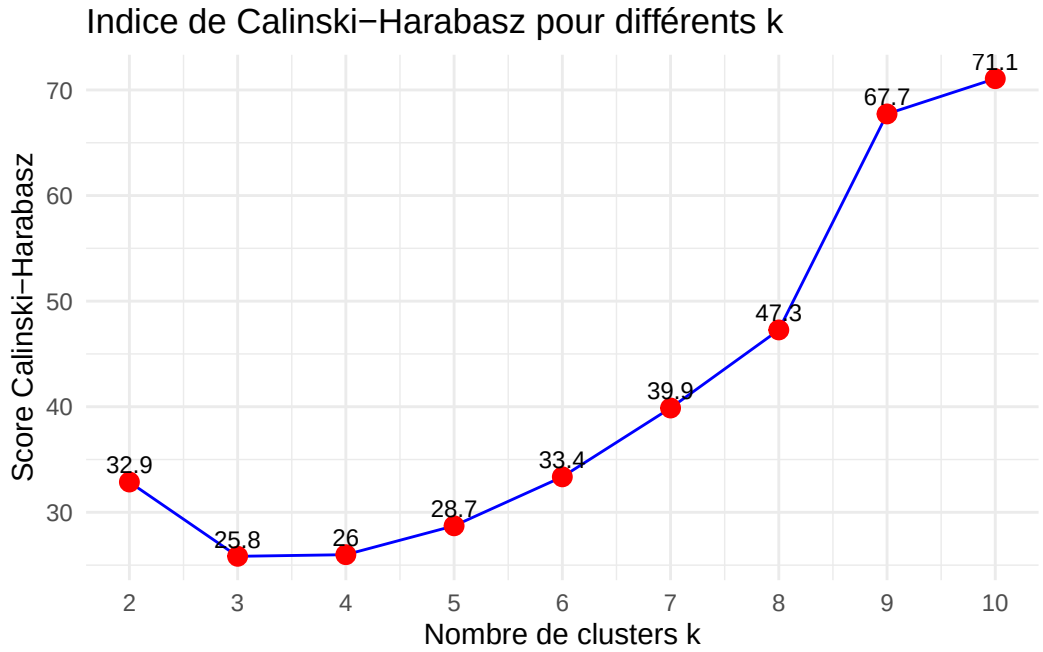
- Ici, on retiendra les 4 premières composantes principales.

5. Réalisons une CAH sur les données en ne retenant que les 4 premières composantes principales :



Faisons la consolidation par k-means :





- La première méthode, basée sur la méthode de la silhouette, suggère un nombre optimal $k = 2$ clusters. Le score 0.764 est convaincant.
- Cependant, couper notre arbre en 2 branches impliquerait d'avoir une classe contenant seulement l'individu 30 : ce qui ne serait pas du tout une bonne classification a proprement parler.
- L'indice de Calinski-Harabasz ne montre pas de pic non plus.

7 Travail personnel 7 (AFDM)

Exercice 30. Données `tennismen`

Nous disposons des données `tennismen` comprenant à la fois des variables quantitatives et des variables qualitatives. Les données étant mixtes, nous allons réaliser une Analyse Factorielle de Données Mixtes (AFDM) sur le jeu `tennismen`.

1. Chargeons les données. Affichons la structure et les premières observations :

Table 7.1 – Structure des données

Variable	Type	Valeurs
Joueur	chr	: chr [1 :20] ”Agassi” ”Becker” ”Borg” ”Connors” ...
Taille	num	: num [1 :20] 180 191 180 178 185 188 187 185 190 190 ...
Lateralite	chr	: chr [1 :20] ”droitier” ”droitier” ”droitier” ”gaucher” ...
MainsRevers	chr	: chr [1 :20] ”deux” ”une” ”deux” ”deux” ...
Titres	num	: num [1 :20] 60 49 64 109 23 79 41 103 26 20 ...
Finales	num	: num [1 :20] 30 28 25 52 13 34 36 54 20 9 ...
TitresGC	num	: num [1 :20] 8 6 11 8 4 17 6 20 2 3 ...
RolandGarros	chr	: chr [1 :20] ”vainqueur” ”demi” ”vainqueur” ”demi” ...
BestClassDouble	num	$ BestClassDouble : num [1 :20] 123 6 890 370 20 114 1 24 4 38 ...

Table 7.2 – Aperçu des données `tennismen`

Joueur	Taille	Lateralite	MainsRevers	Titres	Finales	TitresGC	RolandGarros	BestClassDouble
Agassi	180	droitier	deux	60	30	8	vainqueur	123
Becker	191	droitier	une	49	28	6	demi	6
Borg	180	droitier	deux	64	25	11	vainqueur	890
Connors	178	gaucher	deux	109	52	8	demi	370
Courier	185	droitier	deux	23	13	4	vainqueur	20
Djokovic	188	droitier	deux	79	34	17	vainqueur	114

2. Identifions les variables quantitatives et qualitatives :

```
Variables quantitatives:  Taille Titres Finales TitresGC BestClassDouble

Variables qualitatives:  Joueur Lateralite MainsRevers RolandGarros
```

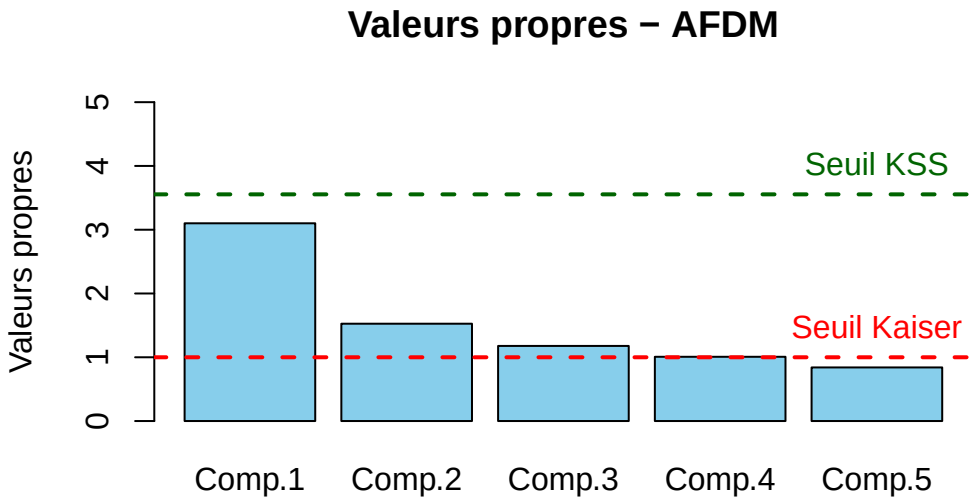
3. Réalisons l’AFDM en utilisant la fonction `FAMD` du package `FactoMineR` :

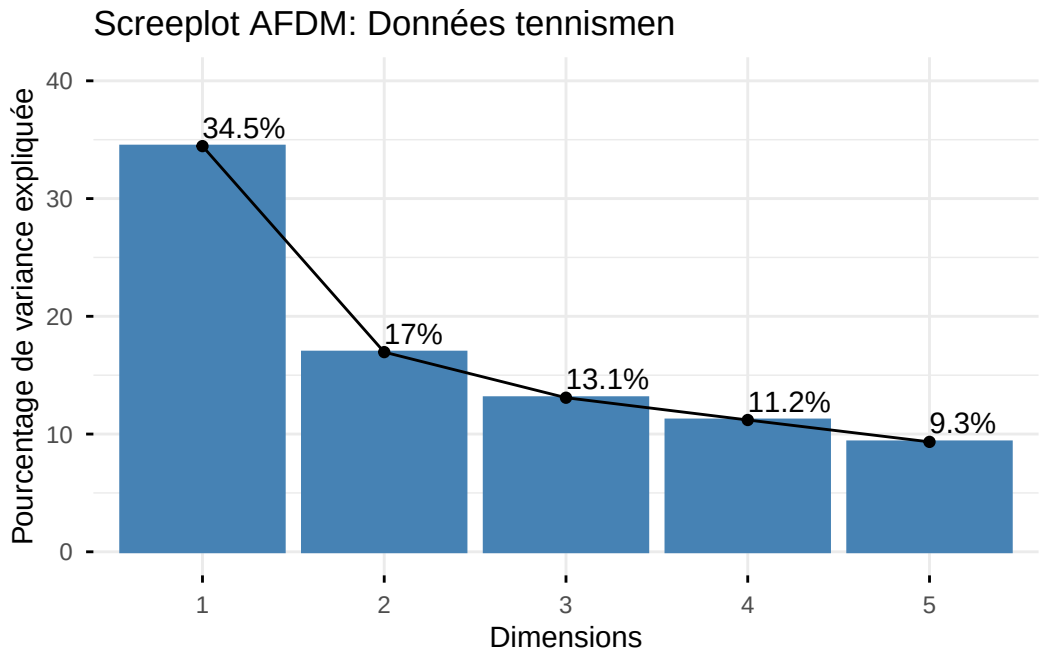
Table 7.3 – Composantes principales de l’AFDM des données `tennismen`

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.1008851	34.45428	34.45428
comp 2	1.5262313	16.95813	51.41240
comp 3	1.1778058	13.08673	64.49914
comp 4	1.0075441	11.19493	75.69407
comp 5	0.8403534	9.33726	85.03133

4. Visualisons les composantes principales et leur pourcentage de variance expliquée :

```
Seuil de Kaiser-Guttman 1

Seuil de Karlis-Saporta-Spinaki 3.554665
```

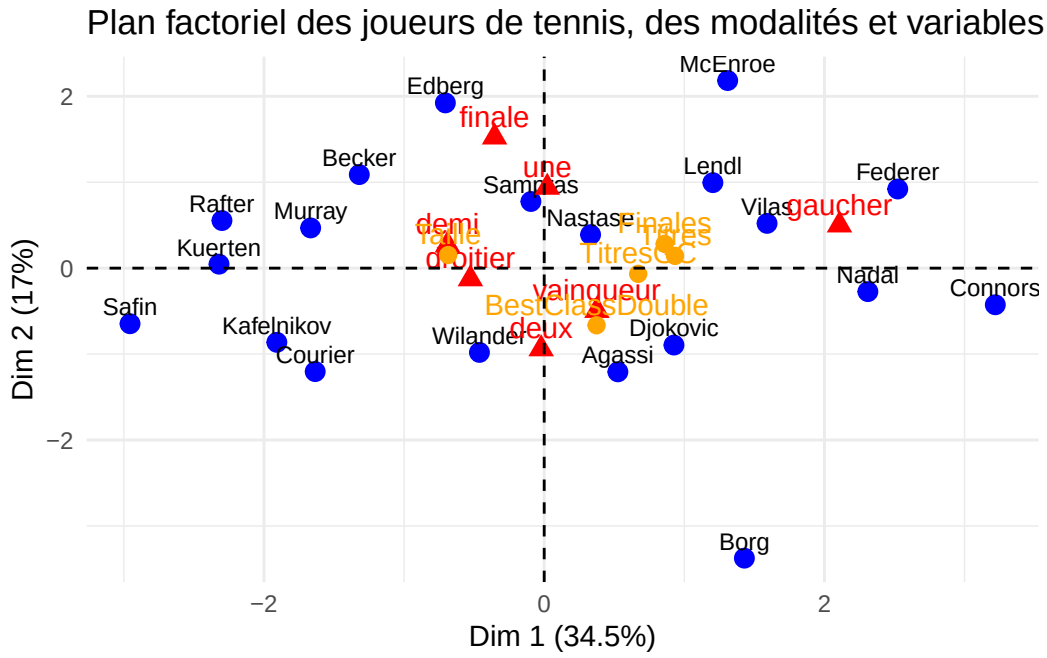




- Comme vous pouvez le remarquer, le seuil de de Karlis−Saporta−Spinaki est trop restrictif.

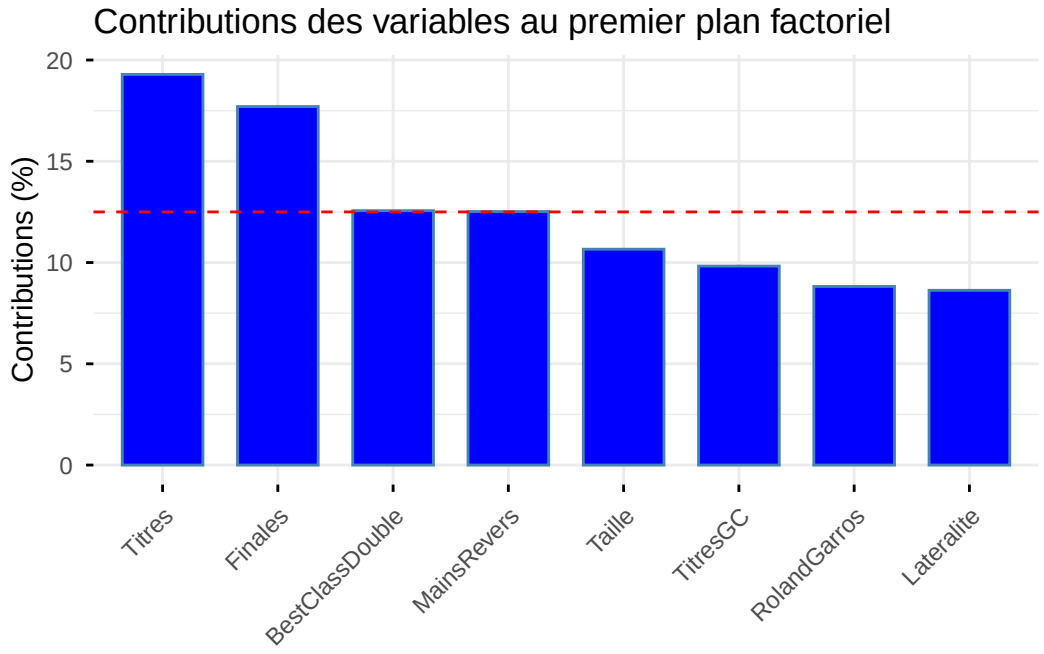
- Par conséquent, selon la Règle de Kaiser–Guttman, on devrait retenir les 4 premières composantes qui ont toutes une valeur propre supérieure à 1 et qui, ensemble, expliquent 75.69% de l'information totale.

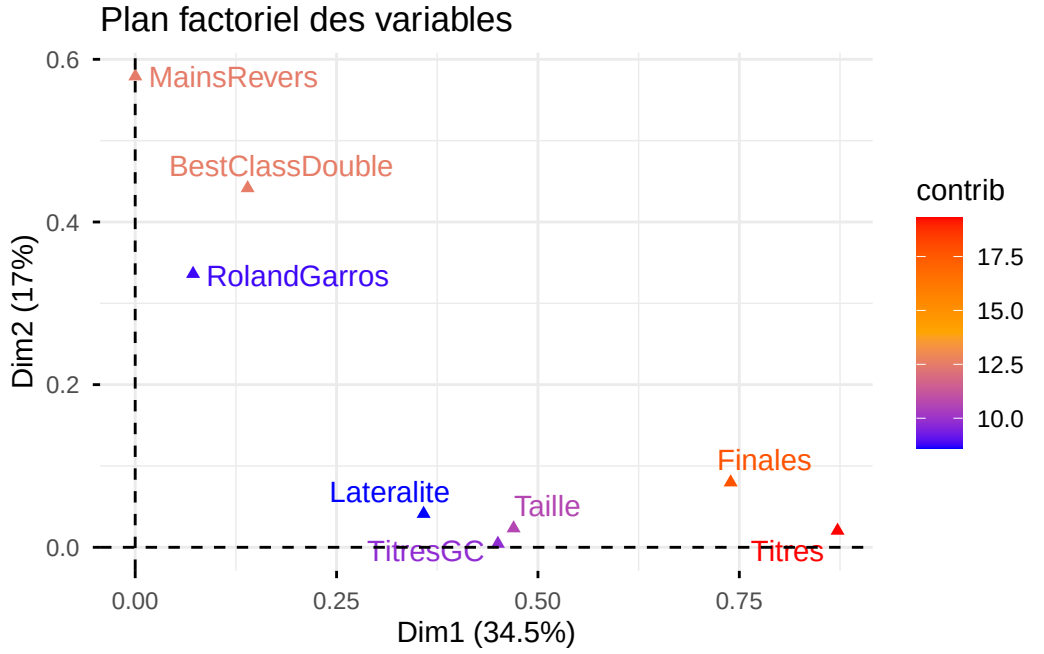
5. Visualisons maintenant les joueurs de tennis sur le premier plan factoriel :



- On peut dire que :
  - L'axe 2 oppose les joueurs de latéralité `gaucher` des individus de latéralité `droitier`.
  - Ce même axe oppose les joueurs qui sont `vainqueurs` de Roland Garros de ceux qui arrivent en `finale` ou en `demi-finale`.
  - L'axe 2 oppose aussi les joueurs qui tiennent la raquette avec `deux mains` de ceux qui la tiennent qu'avec `une main`.
  - Le joueur `Borg` parait atypique vu son éloignement des autres joueurs.
  - Les joueurs de tennis `Federer`, `Nadal`, `Vilas`, `Lendi` et `Connors` sont tous des gauchers.
  - Les joueurs qui sont `droitiers` et qui tiennent la raquette avec `deux mains` comme `Djokovic`, `Agassi` et `Wilander`, sont fortement associés à la modalité `vainqueurs` et aux nombres de `titres Grand Chelem`.
  - Le joueur `Nastase` est fortement associé aux nombres de Finales, aux nombres de Titres.
  - Les gauchers finissent rarement vainqueurs du Roland Garros.
  - Les gauchers comme `Lendl` et `Vilas` ont joué beaucoup de `Finales`.

6. Visualisons maintenant les variables sur le premier plan factoriel :





- Les variables qui contribuent le plus à la dimension 1 sont `Finales` et `Titres`.
- Celles qui contribuent le plus à la dimension 2 sont `MainsRevers` et `BestClassDouble`.

8 Projet personnel

1. Chargeons les données `départements` et faisons un résumé descriptif standard, tout en affichant les premières observations :

Table 8.1 – Structure des données

Variable	Type	Valeurs
TXCR	num	12.19 0.89 -2.63 10 7.82 …
ETRA	num	0.095 0.035 0.031 0.046 0.037 0.091 0.039 0.056 0.055 0.057 …
URBR	num	0.339 0.358 0.464 0.19 0.294 0.945 0.281 0.324 0 0.424 …
JEUN	num	0.285 0.289 0.227 0.235 0.25 0.21 0.252 0.292 0.215 0.27 …
AGE	num	0.133 0.143 0.204 0.19 0.173 0.217 0.184 0.138 0.228 0.155 …
CHOM	num	0.07 0.133 0.135 0.123 0.089 0.119 0.105 0.144 0.12 0.113 …
AGRI	num	0.045 0.049 0.086 0.065 0.072 0.012 0.085 0.065 0.09 0.068 …
ARTI	num	0.084 0.07 0.095 0.127 0.117 0.124 0.106 0.069 0.103 0.067 …
CADR	num	0.088 0.071 0.07 0.09 0.078 0.127 0.073 0.065 0.073 0.067 …
EMPL	num	0.24 0.24 0.262 0.261 0.302 0.316 0.224 0.24 0.261 0.227 …
OUVR	num	0.347 0.4 0.311 0.256 0.23 0.227 0.329 0.386 0.298 0.401 …
PROF	num	0.195 0.171 0.176 0.2 0.201 0.194 0.184 0.176 0.175 0.17 …
FISC	num	2772 2854 3160 4033 3365 …
CRIM	num	38.7 52.2 39 57.7 49.2 …
FE90	num	52.3 56.9 42.7 52.3 53.5 50 48.4 55.6 44.1 51.9 …

Table 8.2 – Résumé statistique de TXCR

Statistique	Valeur
Min.	-5.73
1st Qu.	0.33
Median	2.74
Mean	3.76
3rd Qu.	6.51
Max.	21.87

Table 8.3 – Résumé statistique de ETRA

Statistique	Valeur
Min.	0.01
1st Qu.	0.03
Median	0.05
Mean	0.05
3rd Qu.	0.07
Max.	0.19

Table 8.4 – Résumé statistique de URBR

Statistique	Valeur
Min.	0.00
1st Qu.	0.27

Median	0.40
Mean	0.44
3rd Qu.	0.56
Max.	1.00

Table 8.5 – Résumé statistique de JEUN

Statistique	Valeur
Min.	0.19
1st Qu.	0.24
Median	0.26
Mean	0.26
3rd Qu.	0.28
Max.	0.31

Table 8.6 – Résumé statistique de AGE

Statistique	Valeur
Min.	0.09
1st Qu.	0.13
Median	0.16
Mean	0.16
3rd Qu.	0.19
Max.	0.25

Table 8.7 – Résumé statistique de CHOM

Statistique	Valeur
Min.	0.06
1st Qu.	0.09
Median	0.11
Mean	0.11
3rd Qu.	0.12
Max.	0.17

Table 8.8 – Résumé statistique de AGRI

Statistique	Valeur
Min.	0.00
1st Qu.	0.03
Median	0.06
Mean	0.07

3rd Qu.	0.09
Max.	0.22

Table 8.9 – Résumé statistique de ARTI

Statistique	Valeur
Min.	0.05
1st Qu.	0.07
Median	0.08
Mean	0.09
3rd Qu.	0.10
Max.	0.14

Table 8.10 – Résumé statistique de CADR

Statistique	Valeur
Min.	0.05
1st Qu.	0.07
Median	0.08
Mean	0.09
3rd Qu.	0.10
Max.	0.32

Table 8.11 – Résumé statistique de EMPL

Statistique	Valeur
Min.	0.21
1st Qu.	0.24
Median	0.25
Mean	0.26
3rd Qu.	0.27
Max.	0.33

Table 8.12 – Résumé statistique de OUVR

Statistique	Valeur
Min.	0.13
1st Qu.	0.27
Median	0.31
Mean	0.31
3rd Qu.	0.36

Max.	0.41

Table 8.13 – Résumé statistique de PROF

Statistique	Valeur
Min.	0.14
1st Qu.	0.17
Median	0.18
Mean	0.19
3rd Qu.	0.20
Max.	0.25

Table 8.14 – Résumé statistique de FISC

Statistique	Valeur
Min.	2216.90
1st Qu.	2768.80
Median	2978.00
Mean	3110.26
3rd Qu.	3354.30
Max.	5029.70

Table 8.15 – Résumé statistique de CRIM

Statistique	Valeur
Min.	24.60
1st Qu.	36.75
Median	46.20
Mean	52.06
3rd Qu.	62.25
Max.	139.90

Table 8.16 – Résumé statistique de FE90

Statistique	Valeur
Min.	39.50
1st Qu.	47.70
Median	51.40
Mean	50.70
3rd Qu.	53.75
Max.	64.40

Table 8.17 – Aperçu des données `départements`

TXCR	ETRA	URBR	JEUN	AGE	CHOMAGRI	ARTI	CADR	EMPL	OUVR	PROF	FISC	CRIM	FE90
12.19	0.10	0.34	0.28	0.13	0.07	0.04	0.08	0.09	0.24	0.35	0.20	2772.3	38.7	52.3
0.89	0.04	0.36	0.29	0.14	0.13	0.05	0.07	0.07	0.24	0.40	0.17	2854.4	52.2	56.9
- 2.63	0.03	0.46	0.23	0.20	0.14	0.09	0.10	0.07	0.26	0.31	0.18	3159.6	39.0	42.7
10.00	0.05	0.19	0.23	0.19	0.12	0.06	0.13	0.09	0.26	0.26	0.20	4033.1	57.7	52.3
7.82	0.04	0.29	0.25	0.17	0.09	0.07	0.12	0.08	0.30	0.23	0.20	3364.9	49.2	53.5
10.56	0.09	0.94	0.21	0.22	0.12	0.01	0.12	0.13	0.32	0.23	0.19	4335.4	123.6	50.0

- Notre jeu de données comprend 95 observations pour 15 variables toutes quantitatives. Ceci nous suggère déjà que la méthode d'analyse appropriée pour ce jeu de données est l'Analyse des Composantes Principales (ACP).
- Mais, bien avant de réaliser l'ACP, remarquons d'abord que les variables `FISC`, `CRIM` et `FE90` prennent de très grandes valeurs. Ceci suggère la nécessité d'une standardisation des données.

2. Standardisons nos données :

Table 8.18 – Aperçu des données `départements` standardisées

TXCR	ETRA	URBR	JEUN	AGE	CHOMAGRI	ARTI	CADR	EMPL	OUVR	PROF	FISC	CRIM	FE90
1.7	1.3	-0.4	1.0	-0.8	-1.7	-0.5	-0.1	-0.1	-0.7	0.7	0.3	-0.6	-0.6	0.3
-0.6	-0.5	-0.3	1.1	-0.6	0.9	-0.4	-0.8	-0.5	-0.7	1.6	-0.6	-0.5	0.0	1.3
-1.3	-0.6	0.1	-1.2	1.2	1.0	0.3	0.5	-0.5	0.3	0.0	-0.4	0.1	-0.6	-1.7
1.3	-0.2	-1.1	-0.9	0.8	0.5	-0.1	2.1	-0.1	0.2	-0.9	0.5	1.7	0.3	0.3
0.8	-0.4	-0.6	-0.3	0.3	-0.9	0.0	1.6	-0.3	1.9	-1.4	0.6	0.5	-0.1	0.6
1.4	1.2	2.2	-1.8	1.5	0.3	-1.2	2.0	0.9	2.5	-1.5	0.3	2.3	3.4	-0.1

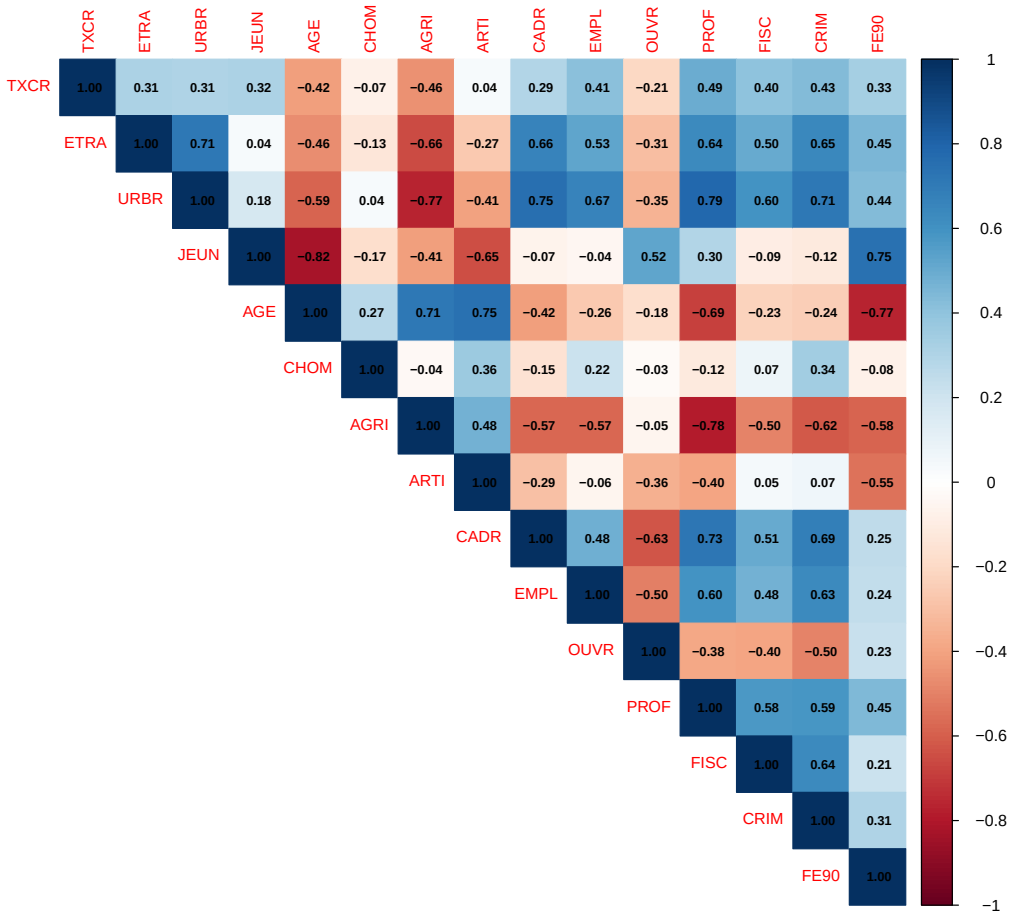
3. Ajoutons les noms de tous les départements comme indices des lignes.

Notons que dans le jeu de données, le nom du 20^{me} département a été omis alors que les données correspondantes ont été bien enregistrées. On le notera par `Dept20`.

Table 8.19 – Aperçu des données `départements` avec les noms

	TXCR	ETRA	URBR	JEUN	AGE	CHOM	AGRI	ARTI	CADR	EMPL
Ain	1.72	1.31	-0.42	0.96	-0.84	-1.66	-0.50	-0.10	-0.10	-0.67
Aisne	-0.58	-0.49	-0.34	1.10	-0.55	0.90	-0.42	-0.82	-0.52	-0.67
Allier	-1.30	-0.61	0.12	-1.17	1.17	0.98	0.32	0.47	-0.55	0.25
Alpes Haute Provence	1.27	-0.16	-1.06	-0.87	0.77	0.50	-0.10	2.11	-0.05	0.21
Hautes Alpes	0.83	-0.43	-0.62	-0.32	0.29	-0.89	0.04	1.60	-0.35	1.93
Alpes Maritimes	1.39	1.19	2.19	-1.79	1.53	0.33	-1.16	1.96	0.86	2.52

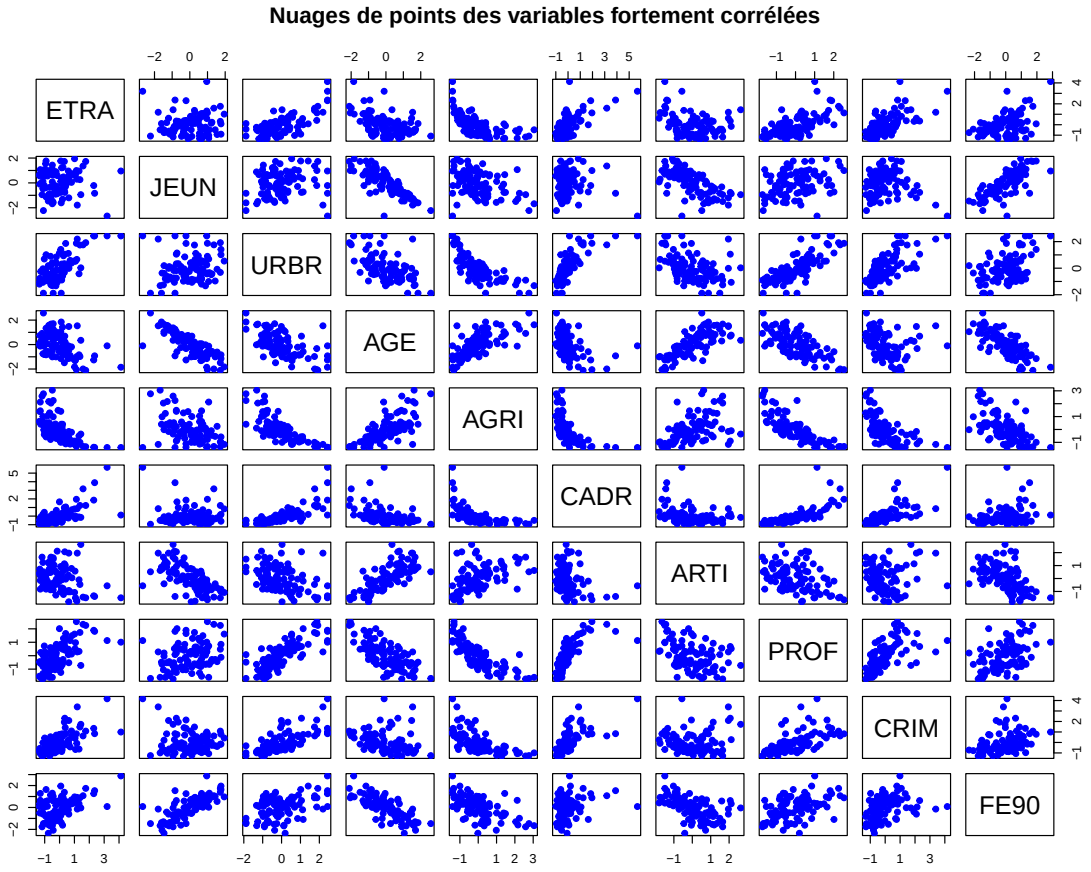
4. Analysons les corrélations entre les variables socioéconomiques :



On peut remarquer que certaines variables sont fortement liées et de manière logique :

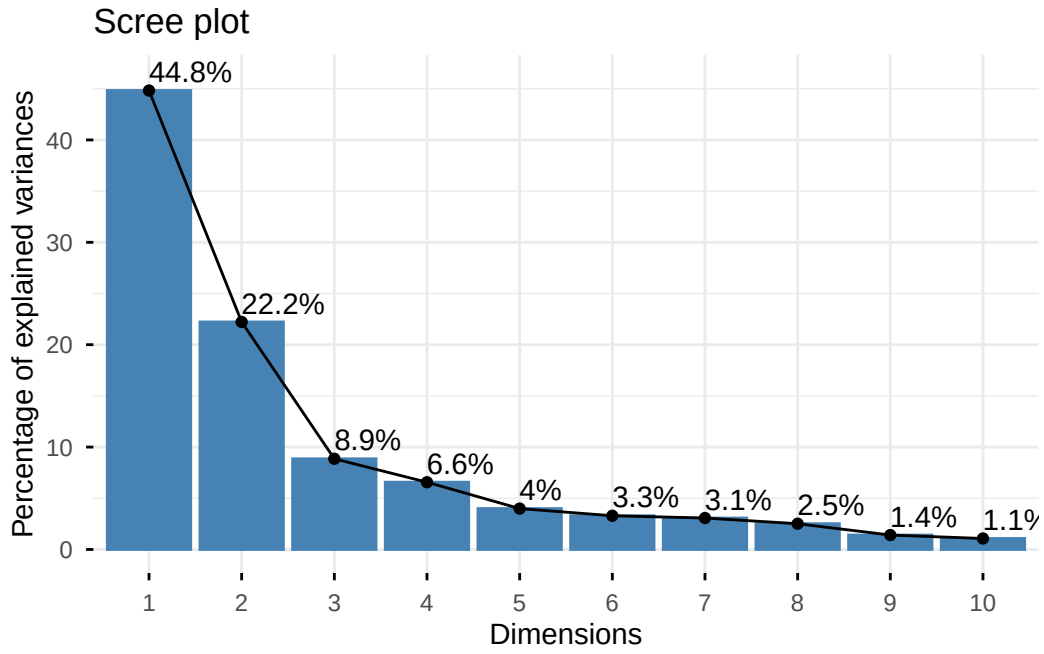
- Les variables AGE et JEUN sont fortement corrélées négativement : ce qui est cohérent puisque plus il y a de jeunes, moins il y a de personnes âgées.
- Les variables URBR (urbanisation) et AGRI (agriculteur) sont aussi fortement corrélées négativement : ce qui est encore logique géographiquement puisqu'en pratique, les agriculteurs ne vivent pas dans les milieux urbains.
- Par ailleurs, les variablesPROF et URBR , tout comme CADR et URBRsont positivement corrélées : c'est encore logique car dans les zones urbaines, il y a généralement plus de cadres supérieurs et plus de professions intermédiaires.
- Les variables CRIM et URBR sont aussi fortement corrélées positivement : ce qui suggère logiquement que le taux de criminalité est plus élevé dan les zones urbaines.
- Et pour finir, les variables AGE et FE90 sont fortement corrélées négativement : ce qui suggère logiquement que le taux de fécondité devient plus faible quand la part de personnes âgées augmente.
- A l'inverse, les variables JEUN et FE90 sont, en effet, fortement corrélées positivement car plus il y a de jeunes, plus il y a de naissances et donc, plus le taux de fécondité augmente.

5. Visualisons les nuages de points par paire pour mettre en évidence ces relations :

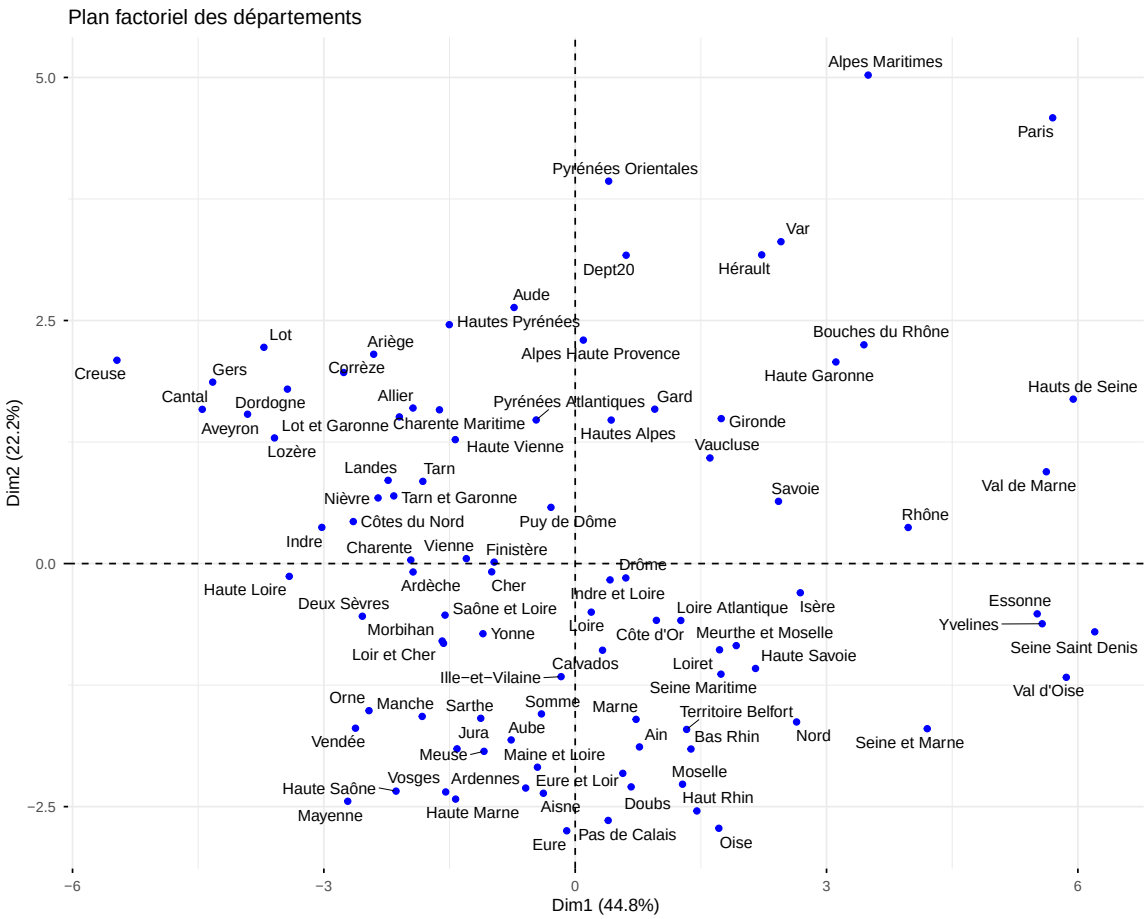


- Les nuages de points confirment, en effet, l'existence de corrélations fortement positives ou fortement négatives entre certaines de nos variables socioéconomiques. Ces relations mettent en évidence des redondances dans l'information contenue dans le jeu de données.
- D'où la nécessité de réaliser une Analyse des Composantes Principales sur notre jeu de données afin de réduire la dimensionalité, de synthétiser l'information et de limiter l'effet de ces corrélations.

6. Réalisons donc une ACP sur notre jeu de données `départements` standardisées :

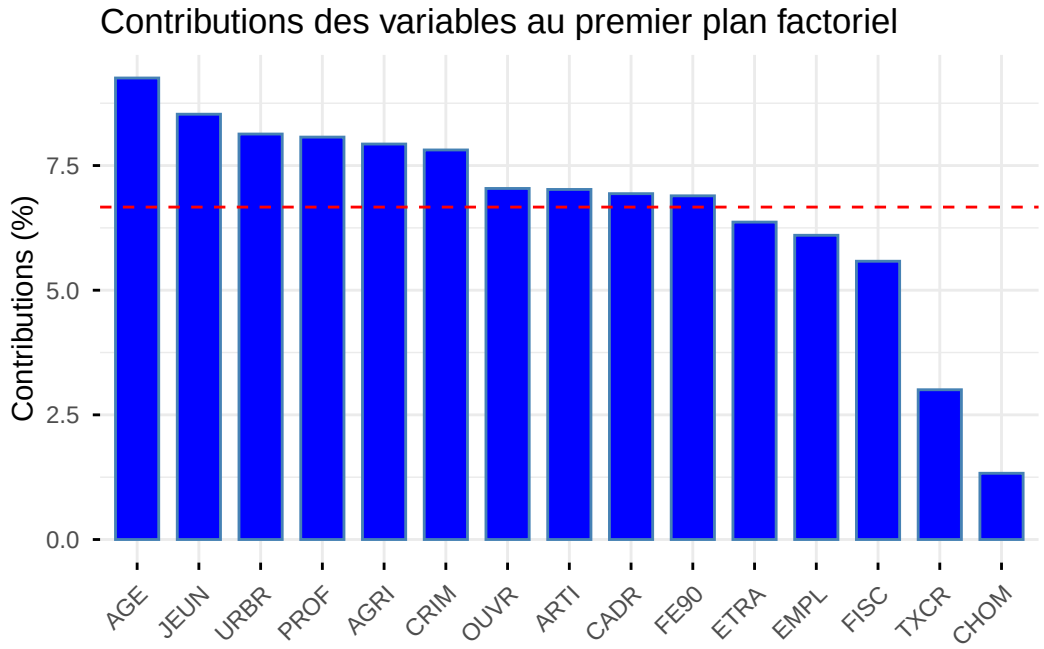


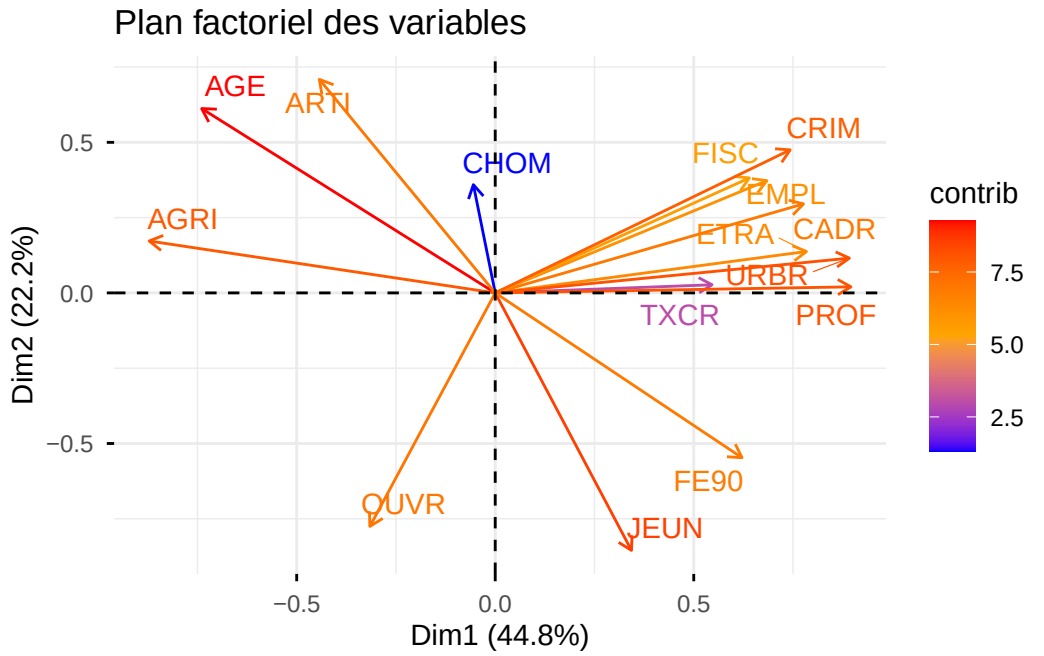
- Pour nos analyses ultérieures, nous retiendrons les **3 premières composantes principales**, qui expliquent ensemble 75,9% de la variance totale.
- Chacune de ces composantes possède une valeur propre supérieure à 1, ce qui signifie qu'elle explique une part de variance supérieure à celle d'une variable d'origine.
- La méthode du coude (critère de Kaiser), appliquée sur le screeplot des valeurs propres, suggère également que les 3 premières composantes suffisent pour expliquer l'essentiel de l'information contenue dans les données.



En analysant le plan factoriel des individus, on peut dire que :

- Les départements comme `Gers`, `Dordogne`, `Cantal`, `Lot` et `Aveyron` sont très proches et donc, ils ont des profils socio-économiques similaires.
- De même, d'autres départements comme `Essonne`, `Yvelines`, `Seine Saint Denis` et `Val d'Oise` ont aussi des profils socio-économiques similaires.
- Par ailleurs, les départements `Paris` et `Alpes Maritimes` présentent un profil socio-économique atypique par rapport aux autres départements, le reste formant un nuage relativement rond.

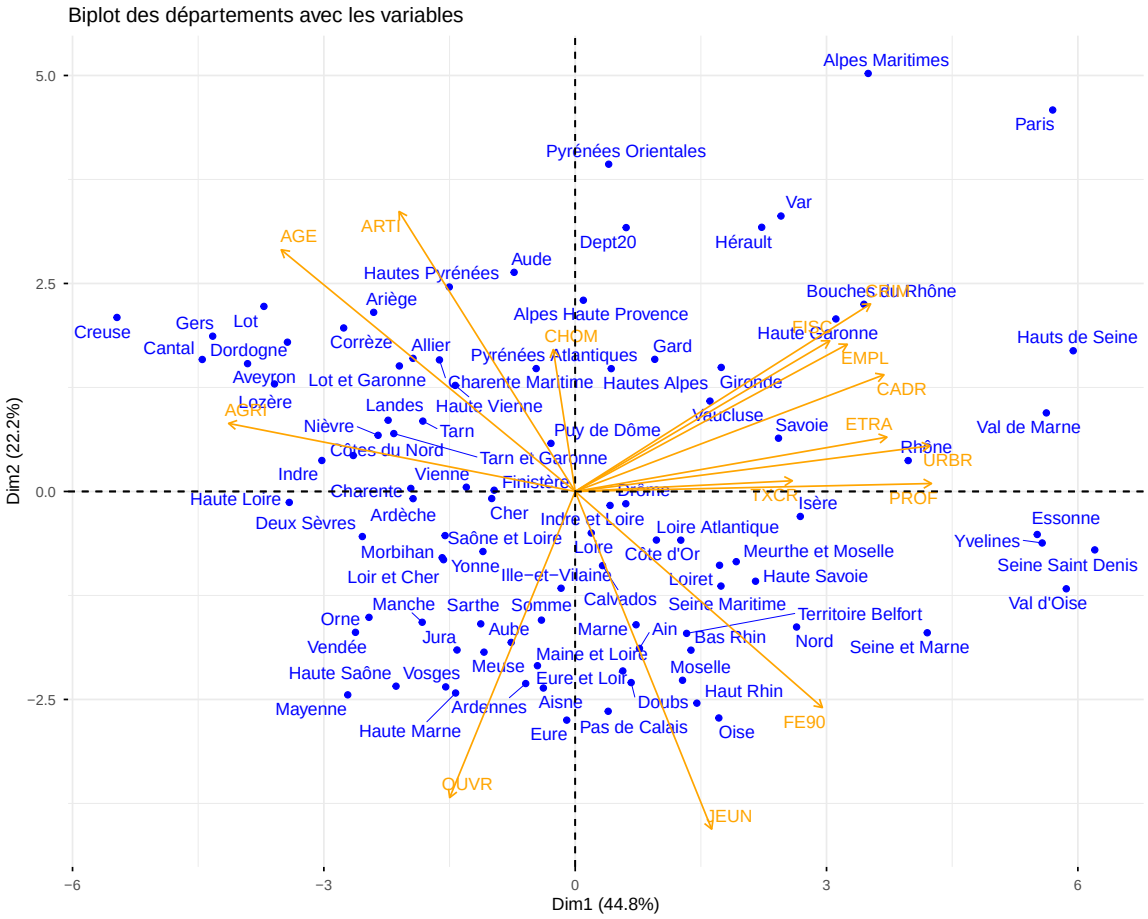




On peut vite remarquer que :

- Les 6 variables les plus contributives au premier plan factoriel sont les variables `AGE`, `JEUN`, `URBR`, `PROF`, `AGRI` et `CRIM` et celle qui contribue le moins est `CHOM`.
- Les variables `FISC`, `CRIM`, `EMPL`, `CADR`, `ETRA`, `URBR`, `PROF` et `TXCR` pointent globalement vers la même direction. Cela indique que ces variables sont fortement corrélées positivement et donc, les départements qui présentent un taux élevé pour l'une tendent également à présenter des valeurs élevées pour les autres.

- Par ailleurs, les variables `AGE`, `ARTI` et `CHOM` pointent plutôt en direction opposée par rapport à `JEUN` et `FE90` : ces relations ont, d'aileurs, été mises en évidence un peu plus tôt dans le rapport.

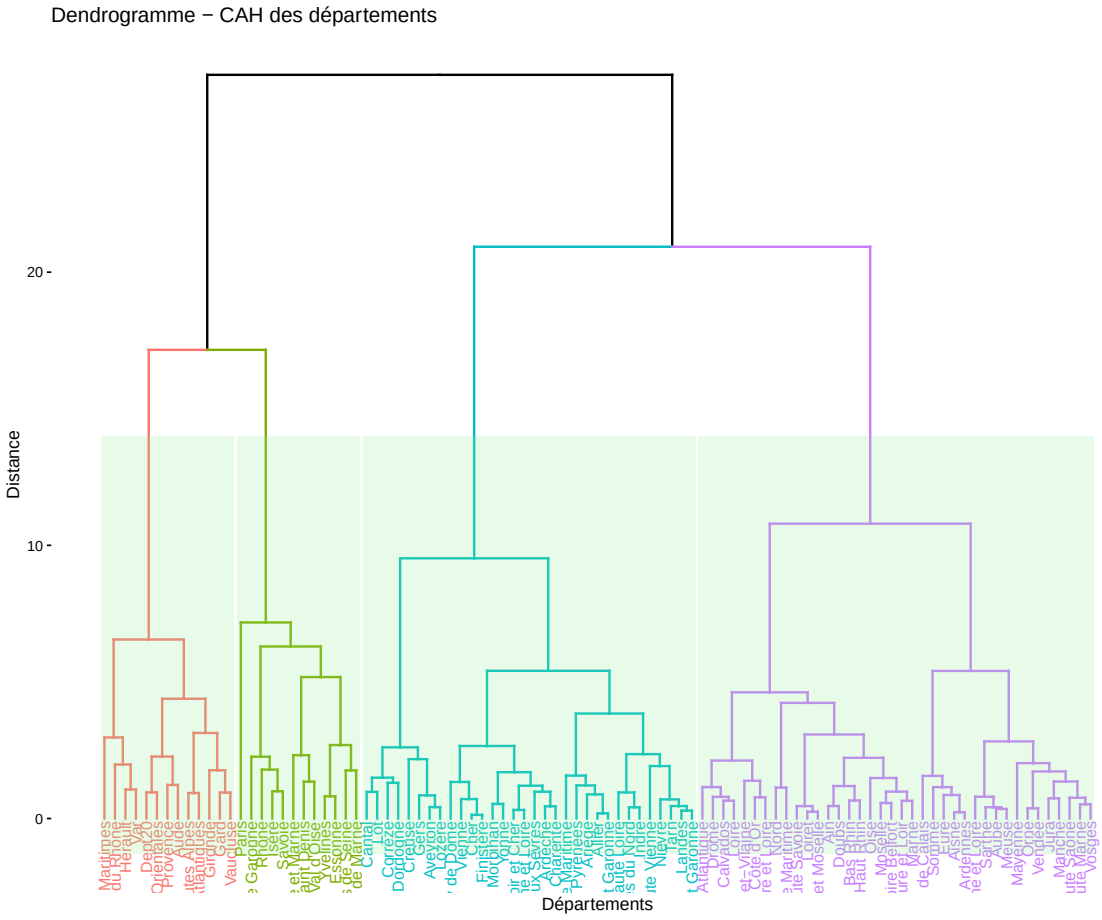


- Comme nous l'avions souligné précédemment, les départements tels que `Haut de Seine`, `Val de Marne`, `Bouche du Rhone`, `Rhone` et `Savoie` et `Isère`, qui présentent un taux élevé d'urbanisation, tendent également à présenter des taux élevés d'emploi, de criminalité et de fiscalité locale ainsi que de croissance démographique plus ou moins marquée. Ces départements se distinguent ainsi par un profil socio-économique plus marqué en termes de population active qualifiée, urbanisation, fiscalité et criminalité.
- Ces départements présentent aussi des parts élevés d'employés, de cadres supérieurs et d'étrangers, ces derniers étant souvent attirés par des opportunités d'emploi ou le tourisme.
- D'autres départements, comme `Pyrénées Atlantiques`, `Alpes Hautes Provences`, `Aude` et `Hautes Alpes`, seraient caractérisés par un fort taux de chômage.
- Et, par ailleurs, d'autres comme `Marne`, `Doubs`, `Pas de Calais` et `Ain` sont caractérisés par une part élevée de jeunes.

En gros, l'axe 1 oppose les départements les plus urbanisés aux départements les plus ruraux tandis que l'axe 2 oppose les départements majoritairement jeunes des départements majoritairement vieux.

On pourrait même dire que l'axe 1 oppose les départements dont l'économie est fortement influencée par le secteur primaire de ceux dont l'économie est surtout marquée par les secteurs tertiaires et quaternaires.

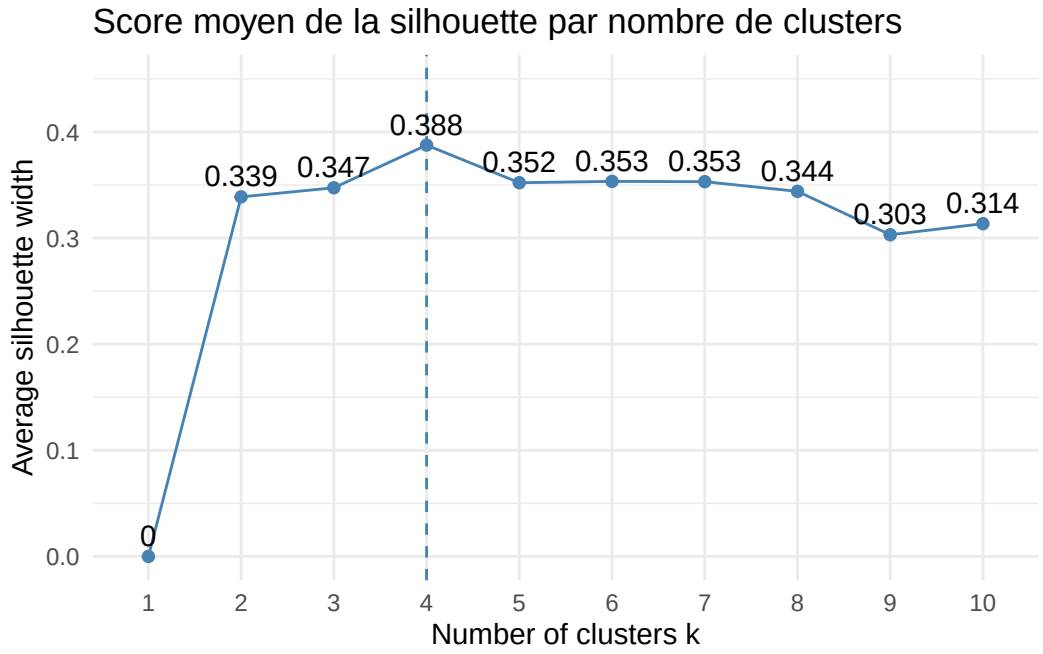
7. Maintenant que nous avons réalisé une ACP sur nos données et choisi de retenir les 3 premières composantes principales, procédons à une **Classification Ascendante Hiérarchique** afin de regrouper les départements présentant des profils socio-économiques similaires :



Sans surprise, on remarque que :

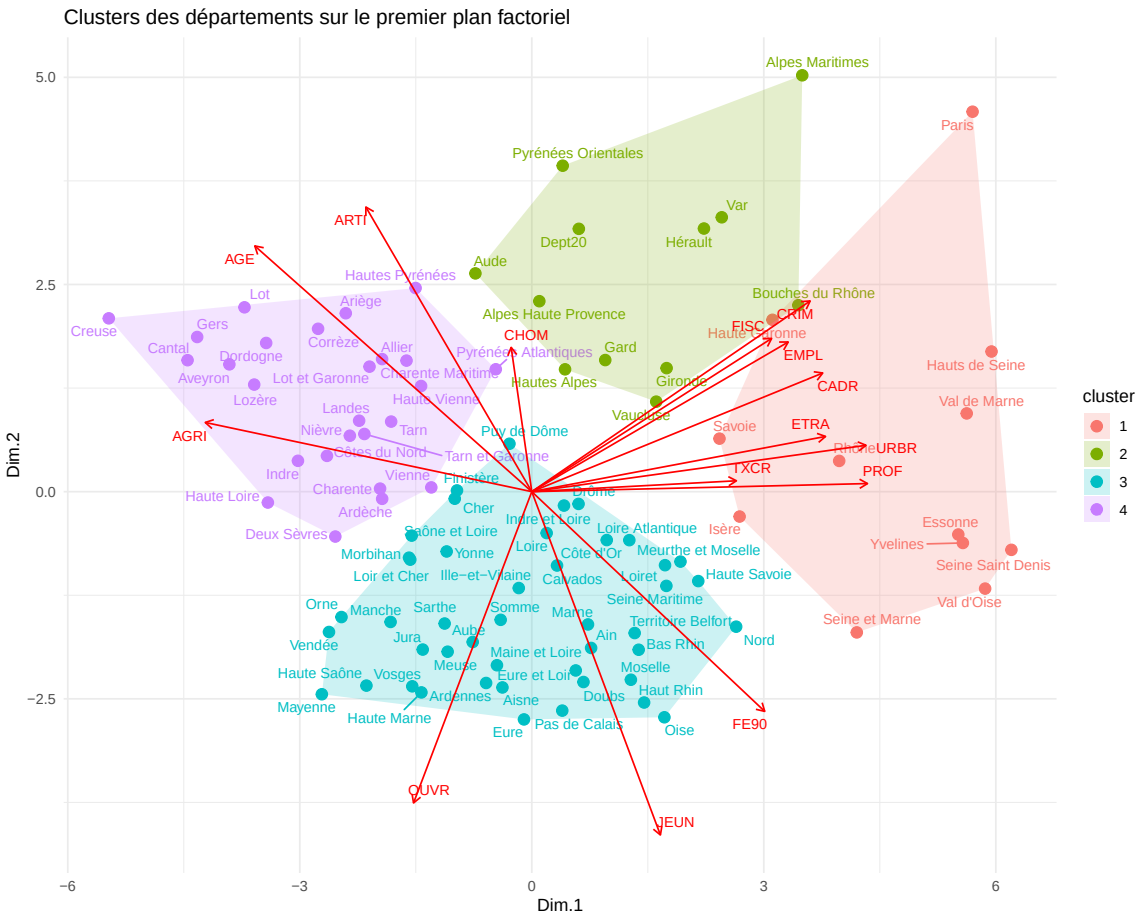
- Les départements comme `Paris`, `Rhône`, `Isère`, `Savoie`, `Val de Marne` et `Hauts de Seine` se retrouvent dans le même groupe.
- En même temps, si on coupait l'arbre plus bas, Paris se serait retouvé tout seul et cela mettrait encore l'emphase sur le fait qu'il s'agit d'un département plutôt atypique.
- Le 20-ème département dont le nom nous est inconnu se retrouve dans le premier groupe avec `Pyrénées Orientales`, `Alpes Hautes Provence`, `Aude` etc...
- Le cluster `bleu` est le plus représenté dans l'échantillon ; il regroupe N individus, ce qui en fait le cluster majoritaire.

8. On va compléter notre analyse par une `consolidation par K-means` et l'évaluation par la méthode de la silhouette pour déterminer le nombre optimal de clusters et la qualité de la classification :



- Ici, pour $k = 4$ clusters, le score est maximal mais reste faible.
- Nous essaierons quand même avec 4 clusters.

9. Réalisons la classification K-means avec **4 clusters** et visualisons les clusters sur le premier plan factoriel :



Nous pouvons observer globalement que :

- Les polygones convexes délimitent correctement les 4 groupes, avec un chevauchement minimal.

Et si on interprétait les classes :

- La classe `rouge` comprend des départements comme `Paris`, `Rhône`, `Isère`, `Savoie`, `Val de Marne` et `Hauts de Seine` et ils sont tous caractérisés par des taux élevés d'urbanisation, d'emploi, de criminalité et de fiscalité locale ainsi que de croissance démographique plus ou moins marquée. Ce sont les départements les **plus urbanisés** et les plus riches en cadres et employés qualifiés.
- La classe `verte` comprend des départements avec des profils **plutôt mixtes** puisque certains comme `Aude`, `Hautes Alpes` sont caractérisés par un fort taux de chômage alors que d'autres comme `Bouches du Rhône`, `Vaucluse` ont des taux d'emplois et de fiscalité plutôt élevés. Tous ces départements se situeraient au sud de la France et auraient une économie fortement marquée par le tourisme. Soulignons la présence à la fois de grandes métropoles ou pôles urbains tels que `Marseille` (Bouches du Rhône), `Montpellier` (Hérault), `Bordeaux` (Gironde) et de zones rurales ou montagneuses à proximité. C'est exactement ce que reflète leur position intermédiaire entre les départements les plus urbanisés et les départements les plus ruraux.
- La classe `violet`, **la plus homogène**, comprend surtout les départements les plus ruraux comme `Ariège`, `Hautes Pyrénées`, `Lozère`, `Corrèze` et `Aveyron`. En termes de population, les habitants sont pour la plupart des vieux. Et ceci explique aussi que leurs parts d'artisans

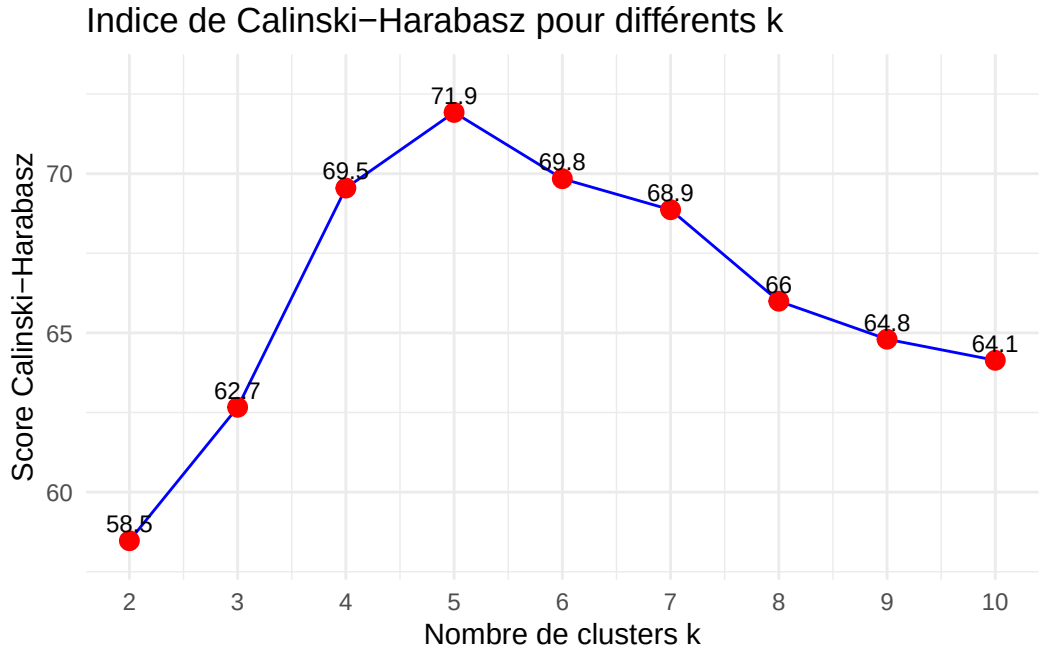
et d'agriculteurs soient les plus élevées. Par ailleurs, on dénote aussi un taux de chômage assez élevé dans ces départements.
- Enfin, nous avons la classe `bleue` avec des départements comme `Pas de Calais`, `Oise`, `Haut Rhin`, `Haute Marne`, `Sarthe` ayant des populations **majoritairement jeunes** à l'opposé de la classe `violet`. Du fait de leur jeunesse, leur taux de fécondité est le plus élevé. Aussi, ces département sont caractérisés par une part d'ouvriers grande ou modérée pour certains.

Dans l'ensemble, cette classification s'avère pertinente et très informative.

Les classes sont bien différenciées, la cohésion interne est satisfaisante, et les profils socio-économiques mis en évidence sont cohérents avec la réalité territoriale française.

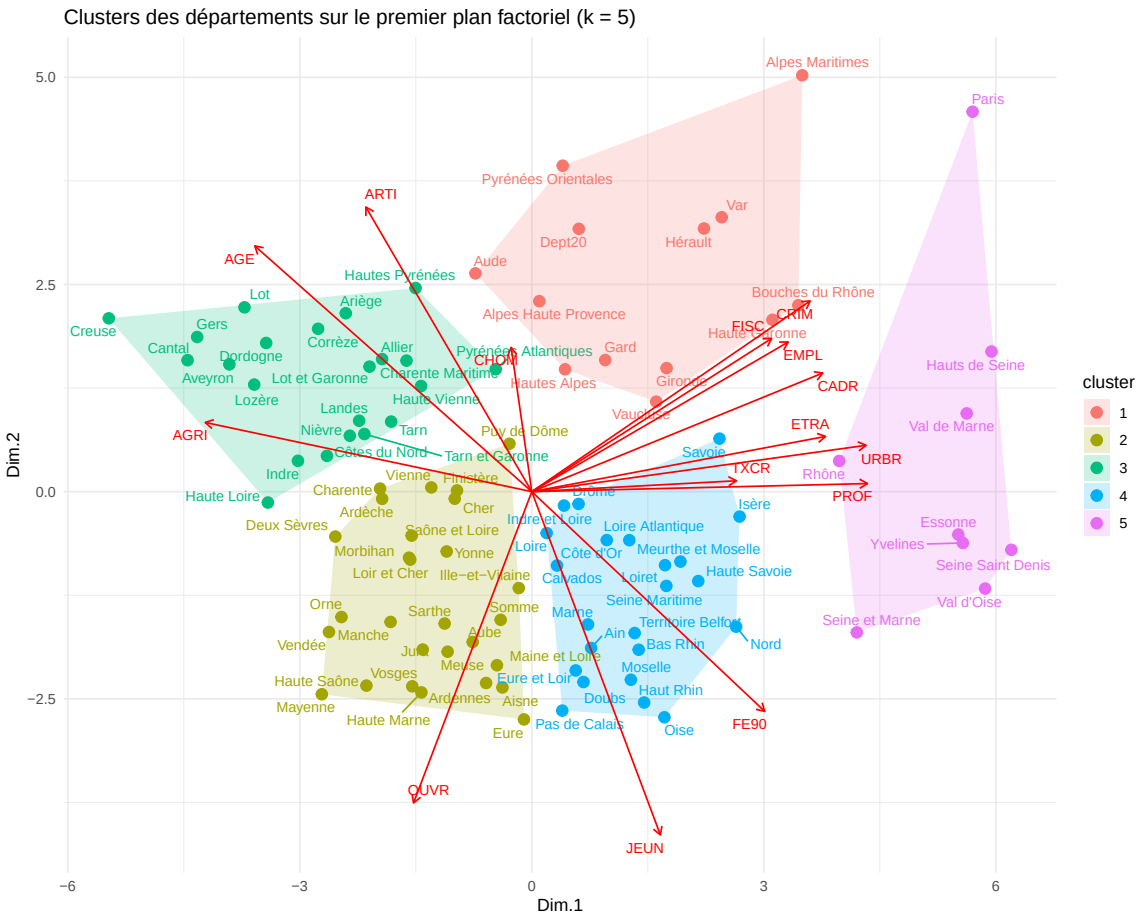
La classe `verte` demeure la plus délicate à interpréter en raison de son hétérogénéité, mais celle-ci s'explique par la diversité économique propre aux départements du sud de la France.

10. On pourrait essayer d'explorer une autre méthode : celle de l'**Indice de Calinski-Harabasz** pour déterminer le nombre optimal de clusters et la qualité de la classification :



- L'indice de Calinski-Harabasz nous suggère une meilleure classification pour un nombre optimal $k = 5$ clusters, où le score est maximal et vaut `71.9`.
- Nous allons donc essayer avec 5 clusters.

11. Réalisons la classification K-means avec **5 clusters** et visualisons les clusters sur le premier plan factoriel :



Par rapport à la classification précédente, plusieurs changements sont notables :

- Cinq classes sont identifiées, ce qui introduit un groupe supplémentaire par rapport à la solution précédente.
- Les interprétations restent globalement les mêmes pour les classes d'avant. Par contre, la nouvelle classe `violet` devient encore plus atypique du fait de éloignement de ces départements par rapport aux autres.
- La classe `bleue` d'avant a été scindée en 2 petites classes : une classe `jaune` et une nouvelle classe `bleue`.
- La nouvelle classe `jaune` regrouperait des départements fortement caractérises par une économie industrielle entretenue par une part élevée d'ouvriers.
- Tandis que la nouvelle classe `bleue` conserve l'ancien profil de départements majoritairement jeunes et plus ou moins industriels.