RAPPORT D'ANALYSE DE DONNEES

LEUCHI ILIAS

Contents

Chapitre 1 : Prise en main et algèbre

Exercice 13

Un échantillon de dossiers d'enfants a été saisi. Ce sont des enfants vus lors d'une visite en 1ère section de maternelle en 1996-1997 dans des écoles de Bordeaux (Gironde, France). L'échantillon est constitué de 152 enfants âgés de 3 ou 4 ans. Les variables sont : le poids de naissance de l'enfant (variable quantitative BWT, exprimée en grammes), l'age de la mère, le poids de la mère lors du dernier cycle menstruel, la "Race" de la mère, le tabagisme durant la grossesse, le nombre d'antécédents de prematurité, l'antécédents d'hypertension, la présence d'irritabilité utérine, le nombre de visites à un médecin durant le premier trimestre de la grossesse, le poids de naissance et le poids de naissance inférieur ou égal à 2500 g.

Table 1: Extrait des données

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

Le tableau 1 est un court extrait du jeu de données. Ici la variable LWT, qui correspond au poids de la mère, est exprimée en livres, nous le modifions donc pour l'avoir en kilogrammes.

Table 2: Extrait des données avec chagement d'unité du poids de la mère

ID	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FVT	BWT	LOW
85	19	82.554	2	0	0	0	1	0	2523	0
86	33	70.307	3	0	0	0	0	3	2551	0
87	20	47.627	1	1	0	0	0	1	2557	0
88	21	48.988	1	1	0	0	1	2	2594	0
89	18	48.534	1	1	0	0	1	0	2600	0
91	21	56.245	3	0	0	0	0	0	2622	0

Nous obtenons le jeu de données du tableau 2.

Réalisons maintenant quelque tri à plats avec ces données.

Table 3: Tri à plat de l'âge de la mère

classes	Effectifs
(14,20.2]	69
(20.2,26.4]	74
(26.4,32.6]	37
(32.6,38.8]	8
(38.8,45]	1

Avec l'âge de la mère, on remarque dans le tableau 3 que 74 des enfants de nos données ont une mère agée entre 20 et 26 ans.

Table 4: Tri à plat du poids de la mère

classes	Effectifs
(36.2,51.7]	60
(51.7,67.1]	87
(67.1,82.6]	26
(82.6,98]	12
(98,113]	4

Le tableau 4 nous apprend que 87 mères des enfants de nos données, ont un poids qui se situe entre 51.7 et 67.1 kg.

Table 5: Tri à plat de la race de la mère

Race	Effectifs
Blanche	96
Noir	26
Autre	67

Le tableau 5 nous indique que 96 mères des enfants des données ont une race dite "Blanche".

Table 6: Tri à plat du tabagisme durant la grossesse

Tabagisme	Effectifs
Non	115
Oui	74

Et pour finir 74 enfants de nos données, avaient une mère qui fumé durant sa grossesse, comme nous indique le tableau 6.

Exercice 14

Nous allons créés un jeu de données personnelles, l'objectif sera de voir les manipulations possibles sur des données.

Table 7: Jeu de données

Mort.a	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.deces
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

Le tableau 7 nous illustre le jeu de données. Chaque individu correspond à un acteur ou on retrouve son nom, prénom, son nombre d'années de carrière, son nombre de films, et la date de sa mort avec l'âge.

Table 8: Prénom du jeu de données

Prénom
Michel
André
Jean
Louis
Lino
Jacques

Nous pouvons extraire une colonne en particulier, dans le tableau 8 c'est « prénom » qui est extrait.

Table 9: Données trié par l'âge du décès

	Age.du.décès	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.deces
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
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

Nous modifions ensuite le nom de la colonne « Mort.a » , en « Age.du.décès », comme le montre le tableau 9. Et pour finir, nous pouvons ordonner le jeu de données selon une condition. Ici nous voulons ordonner par « Age.du.décès » croissant, ce qui est fait dans le tableau 9 également.

Exercice 15

Le goût d'un fromage dépend de la concentration de plusieurs composés chimiques, dont : la concentration d'acide acétique (variable X1), la concentration d'hydrogène sulfuré (variable X2), la concentration d'acide lactique (variable X3). Pour 30 types de fromage, on dispose du score moyen attribué par des goûteurs (caractère Y).

Table 10: Extrait du jeu de données fromage

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

Voici un extrait des données représentées dans le tableau 10 .On retrouve bien un total de 30 individus qui correspondent à des types de fromage. Il y a bien les 4 variables toutes quantitatives.

Table 11: Statistiques élémentaires des données fromage

Y	X1	X2	X3
Min. : 0.70	Min. :4.477	Min. : 2.996	Min. :0.860
1st Qu.:13.55	1st Qu.:5.237	1st Qu.: 3.978	1st Qu.:1.250
Median :20.95	Median :5.425	Median : 5.329	Median :1.450
Mean :24.53	Mean :5.498	Mean : 5.942	Mean :1.442
3rd Qu.:36.70	3rd Qu.:5.883	3rd Qu.: 7.575	3rd Qu.:1.667
Max. :57.20	Max. :6.458	Max. :10.199	Max. :2.010

Le tableau 11 nous montre les statistiques élémentaires pour chacune des variables. Par exemple pour Y on trouve une valeur moyenne de 24.53, un minimum de 0.7 et un maximum de 57.20.

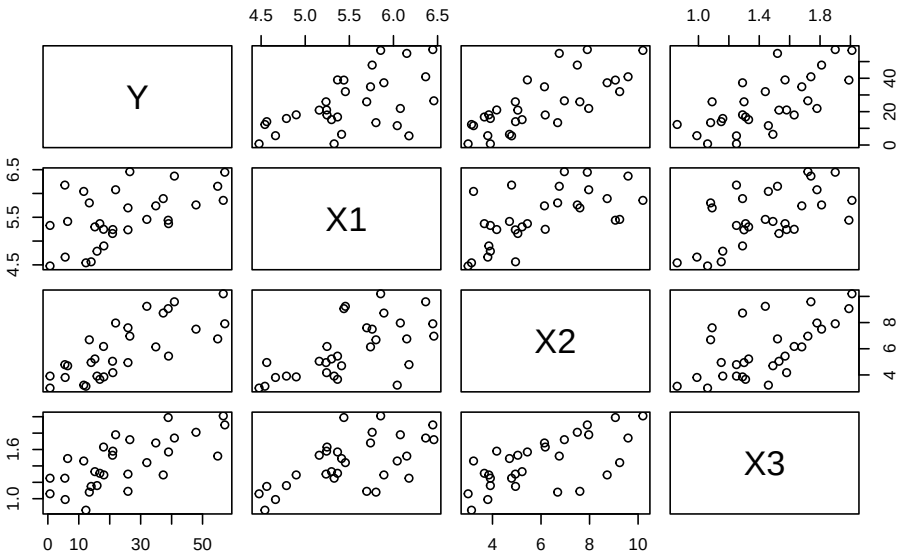

Figure 1: Ozone en fonction des saisons

La figure 1 représente la matrice de nuage de points entre chacune des variables. Ce sont les nuages de points des croisements deux à deux entre chaque variable de nos données.

Table 12: Extrait des données fromage filtrées

	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
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

Nous allons maintenant créer un sous-jeu de données avec les contraintes suivantes : X1 $>$ 5.1 et X3 $<$ 1.77. C'est ce qui est représenté dans le tableau 12.

Table 13: Statistiques élémentaires des données fromage filtrées

Y	X1	X2	X3
Min. : 0.70	Min. :5.159	Min. :3.219	Min. :1.080
1st Qu.:14.30	1st Qu.:5.313	1st Qu.:4.744	1st Qu.:1.295
Median :21.00	Median :5.455	Median :5.438	Median :1.460
Mean :23.52	Mean :5.654	Mean :5.946	Mean :1.435
3rd Qu.:33.45	3rd Qu.:5.968	3rd Qu.:6.857	3rd Qu.:1.575
Max. :54.90	Max. :6.458	Max. :9.588	Max. :1.740

Après ce changement on trouve certaines valeurs différentes des statistiques élémentaires, par exemple la moyenne de Y est maintenant de 23.52. Nous voyons les nouvelles statistiques dans le tableau 13.

Exercice 16

Les données que nous utiliserons sont directement implantées dans R, il s'agit des données «airquality».

Table 14: Extrait 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

Le tableau 14 nous montre un extrait des données. Il s'agit des relevés quotidiens des valeurs de qualité de l'air, du 1er mai 1973 au 30 septembre 1973. Il y a 153 individus pour 6 variables : Ozone taux d'ozone en ppb (parts per billion), Solar.R Rayonnement solaire (langleys), Wind Vitesse du vent (miles par heure) Temp température (degrés Fahrenheit), Month mois (entre 1 et 12), Day jour du mois (entre 1 et 31). Les données ont été obtenues auprès du New York State Department of Conservation (données sur l'ozone) et du National Weather Service (données météorologiques).

Table 15: Statistiques élémentaires des données

Ozone	Solar.R	Wind	Temp
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :56.00
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400	1st Qu.:72.00
Median : 31.50	Median :205.0	Median : 9.700	Median :79.00
Mean : 42.13	Mean :185.9	Mean : 9.958	Mean :77.88
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500	3rd Qu.:85.00
Max. :168.00	Max. :334.0	Max. :20.700	Max. :97.00
NA's :37	NA's :7	NA	NA

Le tableau 15 nous montre les statistiques élémentaires sur nos variables quantitatives, ainsi que les valeurs manquantes. Pour la variable ozone on remarque 37 valeurs manquantes et une moyenne de 42.13.

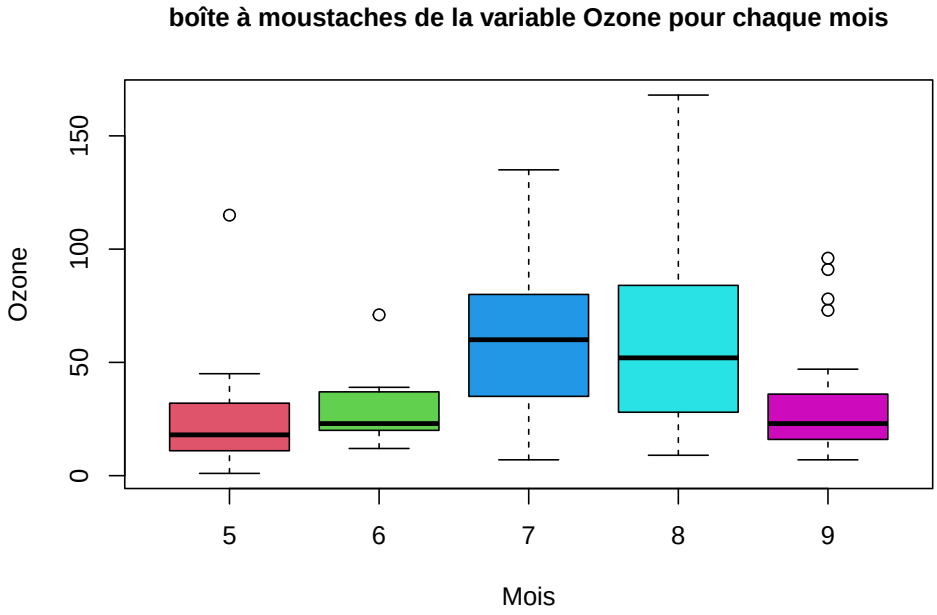

Figure 2: boîte à moustaches de l'Ozone pour chaque mois

On remarque, grâce à la figure 2, des diagrammes à moustache avec une tendance similaire pour les mois 5, 6, et 9, qui ont des valeurs de l'ozone peu élevées, qui varie moins. Alors que pour les mois 7 et 8, les valeurs de l'ozone sont plus fortes et beaucoup plus réparties.

Table 16: Extrait des données avec la saison

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

Pour notre analyse, nous rajoutons une variable saison. Le tableau 16 montre nos données avec cette nouvelle variable.

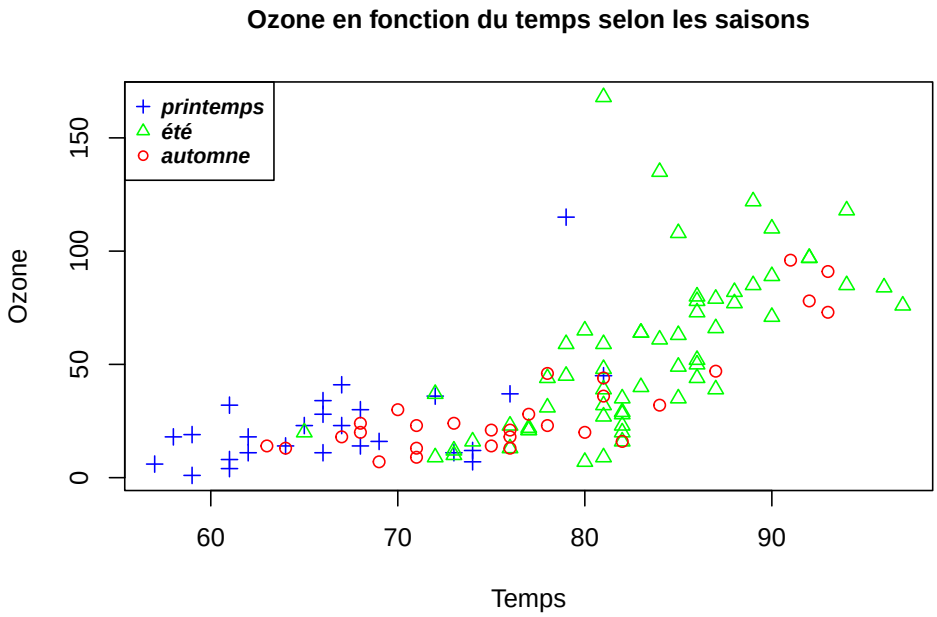

Figure 3: Ozone en fonction des saisons

Avec la figure 3 on remarque qu'il y a une relation positive linéaire croissante entre l'ozone et le temps. Cette relation est présente pour chacune des saisons. Le temps est plus élevé en été et plus faible en hiver. Et comme vu précédemment avec la figure 2, la concentration d'ozone est plus forte en été, qui correspond aux mois 7 et 8.

Exercice 17

Nous nous intéressons à la fonction suivant $yi = 1.7 + 2.1i + ei$, avec i entre 1 et 100, et les ei suivant une lois $N(0, 5\check{s})$.

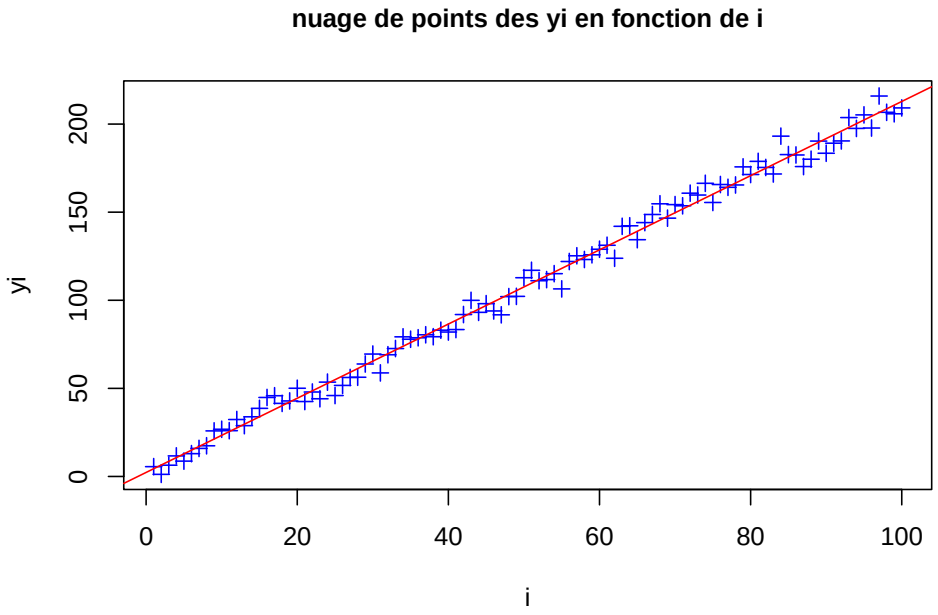

Figure 4: Nuage de points des yi en fonction de i

La figure 4 nous montre le nuage de points généré avec notre fonction, avec la droite de régression. Cette droite semble être un bon ajustement de notre fonction.

Exercice 18

On considère un tableau de contingence obtenu en ventilant 592 femmes suivant la couleur de leurs yeux et la couleur de leurs cheveux.

Table 17: Tableau de contingence du croisement entre la couleur des yeux et des cheveux

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

Le tableau 17 illustre ce tableau de contingence. On apprend par exemple que 94 femmes de nos données sont blondes aux yeux bleus.

Table 18: Matrice des fréquences du croisement entre la couleur des yeux et des cheveux (%)

	blond	brun	châtain	roux
bleu	16	3	14	3
marron	1	11	20	4
noissette	2	3	9	2
vert	3	1	5	2

Et voici la matrice des fréquences de nos données dans le tableau 18. On remarque que 11% des femmes de nos données sont brunes aux yeux marron.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 19: Tableau de contingence avec les marges

	blond	brun	châtain	roux	sum
bleu	94	20	84	17	215
marron	7	68	119	26	220
noissette	10	15	54	14	93
vert	16	5	29	14	64
sum	127	108	286	71	592

Nous pouvons ajouter les marges dans notre tableau de contingence comme le montre le tableau 19. Elles nous informent que 215 femmes de nos données ont les yeux bleus, ou encore que 71 d'entre elles sont rousses.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 20: Distributions conditionnelles des couleurs des cheveux sachant la couleur des yeux

	blond	brun	châtain	roux	sum
bleu	43.72	9.30	39.07	7.91	100
marron	3.18	30.91	54.09	11.82	100
noissette	10.75	16.13	58.06	15.05	100
vert	25.00	7.81	45.31	21.88	100

Le tableau 20 des distributions conditionnelles de la couleur des cheveux, sachant la couleur des yeux, nous apprend par exemple que, 58,06% des femmes aux yeux noisettes ont une couleur de cheveux châtains.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 21: Distributions conditionnelles des couleurs des yeux sachant la couleur des cheveux

	blond	brun	châtain	roux
bleu	74.02	18.52	29.37	23.94
marron	5.51	62.96	41.61	36.62
noissette	7.87	13.89	18.88	19.72
vert	12.60	4.63	10.14	19.72
sum	100.00	100.00	100.00	100.00

Le tableau 21 correspond aux distributions conditionnelles des couleurs des yeux, sachant la couleur des cheveux. On apprend que les femmes brunes ont pour 13.89% d'entre elles les yeux noisette.

Table 22: Matrice des taux de liaisons

	blond	brun	châtain	roux
blond	1.000	-0.241	0.134	-0.155
brun	-0.241	1.000	0.920	0.986
châtain	0.134	0.920	1.000	0.918
roux	-0.155	0.986	0.918	1.000

Le tableau 22 nous apprend les liaisons (variant entre -1 et 1), entre les modalités de la variable couleur des cheveux. Par exemple on voit qu'il y a un lien positif entre la couleur châtain et brun, ce qui signifie qu'il y a une tendance similaire entre ces deux modalités vis-à-vis de la couleur des yeux.

```
##
##  Pearson's Chi-squared test
##
## data:  TEC
## X-squared = 138.29, df = 9, p-value < 2.2e-16
```

On réalise un test sur nos deux variables pour savoir elles sont indépendantes. On remarque une pvaleur très proche de zéro.Ce qui nous permet de rejeter l'hypothèse d'indépendance entre la couleur des yeux et celle des cheveux. Il y a donc un lien entre ces deux variables.

Chapitre 2 : Mesure de la liaison entre une variable et un ensemble de variables

Exercice 19

Nous étudions ici un croisement entre des classes d'âges et des diplômes, pour 90 individus.

Table 23: Tableau de contingence des l'âges croisé avec les diplômes

	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

Le tableau 23 nous donne les effectifs croisés de nos deux variables. On apprend par exemple que dans nos données il y a 40 individus avec un BEPC, et que 15 d'entre eux ont plus de 50 ans.

Table 24: Tableau des fréquences croisées (%)

	BEPC	BAC	Licence	Total
Plus de 50 ans	16.7	13.3	3.3	33.3
Entre 30 et 50 ans	11.1	20.0	4.4	35.6
Moins de 30 ans	16.7	5.6	8.9	31.1
Total	44.4	38.9	16.7	100.0

On peut obtenir les fréquences de notre tableau 23. C'est ce qu'illustre le tableau 24, on voit par exemple que 35.6% de nos individus ont entre 30 et 50 ans, et 20% d'entre eux ont un bac.

Table 25: Profils lignes (%)

	BEPC	BAC	Licence	Total
Plus de 50 ans	50.00	40.00	10.00	100
Entre 30 et 50 ans	31.25	56.25	12.50	100
Moins de 30 ans	53.57	17.86	28.57	100
Total	44.44	38.89	16.67	100

Le tableau 25 nous donne les fréquences sachant la tranche d'âge. On voit que 28.57% des moins de 30 ans ont une licence.

Table 26: Profils collones (%)

	BEPC	BAC	Licence	Total
Plus de 50 ans	37.5	34.29	20.00	33.33
Entre 30 et 50 ans	25.0	51.43	26.67	35.56
Moins de 30 ans	37.5	14.29	53.33	31.11
Total	100.0	100.00	100.00	100.00

Le tableau 26 nous donne les fréquences sachant le diplôme, on apprend que pour ceux ayant un bac, 51.43% ont entre 30 et 50 ans.

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 11.175, df = 9, p-value = 0.2639
```

On cherche à savoir si nos deux variables sont indépendantes. On trouve une pvlaur de 0.2639, une valeur conséquente qui nous permet de conclure sur le non-rejet de H0, et d'en déduire qu'il y a indépendance entre les deux variables.

Exercice 20

Pour une population d'effectifs de taille 1000 on a mesuré les deux variables qualitatives "Couleur des yeux" et "Etat matrimonial".

Table 27: Tableau de contingence de la couleur des yeux et de l'état matrimonial

	Bleu	Brun
Celib	290	410
Marie	110	190

Le tableau 27 nous apprend que parmi nos 1000 individus 290 sont célibataires avec les yeux bleus. Ou encore que 190 sont mariés avec les yeux bruns.

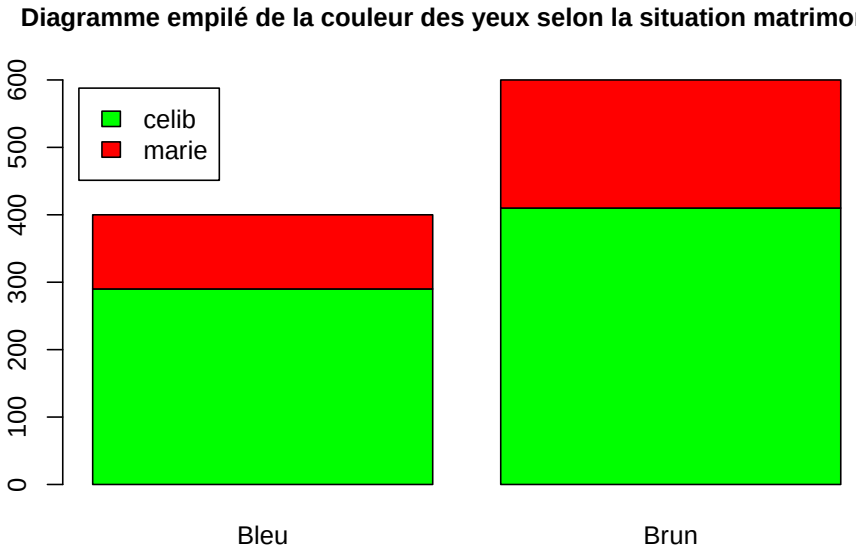

Figure 5: Diagramme empilé de la couleur des yeux selon la situation matrimoniale

On peut rendre graphique le tableau 27. Grâce à la figure 5 on remarque que nous avec moins d'individus aux yeux bleus qu'aux yeux bruns.

Quelque commande R utile sur nos données :

n <- margin.table(tableau) => effectif total

m1 <- margin.table(tableau,1) => lois marginale de l'état matrimonial

m2 <- margin.table(tableau,2) => lois marginale de la couleur des yeux

prop.table(tableau) => tableau de contingence en fréquence

Table 28: Tableau de contingence (%)

	Bleu	Brun
Celib	29	41
Marie	11	19

Voici par exemple le tableau de contingence en pourcentage cette fois. On apprend que 41% des individus sont celibataires au yeux bruns.

Table 29: Tableau des effectifs théoriques

	Bleu	Brun
Celib	280	420
Marie	120	180

Le tableau 29 nous montre les effectifs théoriques, c'est-à-dire les effectifs si nos variables étaient parfaitement indépendantes.

```
## 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
```

La statistique du χ^2 mesure l'écart entre le tableau de contingence et le tableau des effectifs théoriques.

Les résultats du test du χ^2 indiquent une pvaleur supérieur à 0.05 ce qui ne nous permet pas de rejetter H0, l'hypothèse d'indépendance.

```
## 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
```

Si on réalise le test sur le tableau des effectifs théoriques, on trouve une pvaleur de 1. Ce qui prouve que ces données reflète l'indépendance parfaite.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tableau2
## X-squared = 995.84, df = 1, p-value < 2.2e-16
```

quand on réalise le test sur un tableau truqué, où tous les individus aux yeux bleus sont mariés et tous les autres sont célibataires, on trouve une pvaleur très petit. Donc un rejet de H0, il y a une forte dépendance entre les deux variables.

Exercice 21

Pour cet exercice nous utiliserons le jeu de données “cars” directement implanté dans R.

Table 30: Extrait du jeu de données cars

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

Ces données indiquent la vitesse de 50 voitures et les distances nécessaires pour s’arrêter. Notez qu’elles ont été enregistrées dans les années 1920. Nous retrouvons un extrait de ces données dans le tableau 30.

La matrice contient 50 lignes et 2 colonnes. Il y a donc deux variables qui sont “speed” la vitesse en mph, et “dist” la distance d’arrêt en ft.

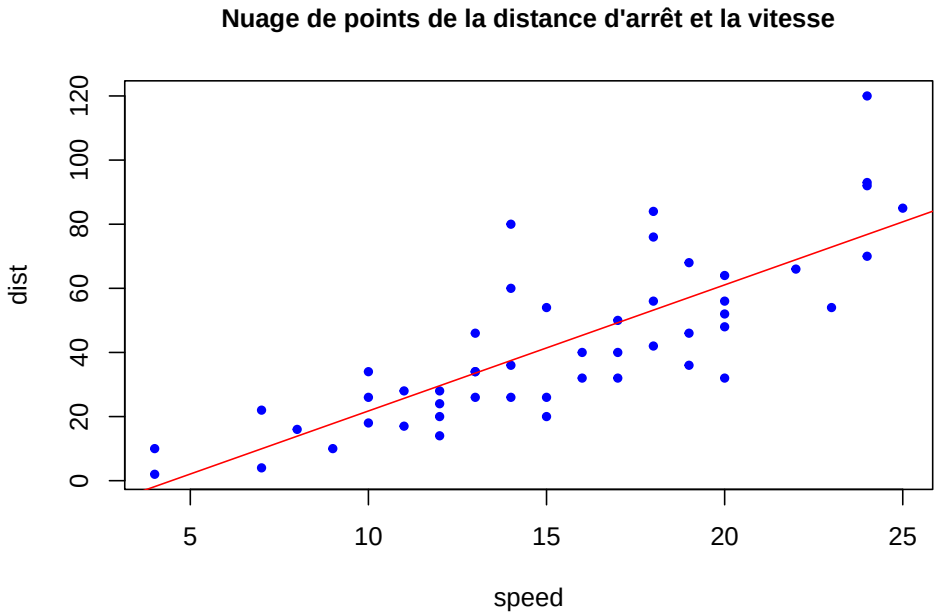

Figure 6: Nuage de points de la distance d’arrêt et la vitesse

Un nuage de points est une bonne représentation entre deux variables quantitatives. La figure est donc adaptée à nos données. Les points semblent liés linéairement de manière positive et croissantes. La droite de régression, en rouge, est celle qui passe le plus près de tous les points.

Le modèle prédit une distance de freinage de 61.07 ft pour une vitesse de 20 mph.

Intervalle de confiance : [55.25 , 66.89]

Intervalle de prédiction : [29.6 , 92.54]

L'exemple cars est adapté à la sélection de modèles.

update() : Va mettre à jour et par défaut réajuster un modèle.

step() : Sélectionnez un modèle basé sur une formule par AIC.

Exercice 22

Nous utiliserons des données extraient d'un recueil issu d'une enquête portant sur une population d'enseignants de collège.

Table 31: Extrait des données

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

Le tableau 31 est un extrait des données que nous utiliserons. Il y a un total de 11 variables, pour 168 individus. La plupart des variables sont explicites. Le salaire est exprimé en euros, l'âge et l'ancienneté en année. Le stress, l'estime de soi et la satisfaction au travail sont mesurés sur des échelles allant de 0 à 50 suivants des techniques appropriées.

Table 32: Résumé statistique du salaire

Salaire
Min. :1200
1st Qu.:1650
Median :1720
Mean :1778
3rd Qu.:1908
Max. :2200

Quand on s'intéresse de près aux salaires dans nos données, grâce au tableau 32, on trouve un minimum de 1200€, un maximum de 2200€, et un salaire médian de 1720€.

Croisement qualitatif vs qualitatif : Sexe et EtatCivil

Essayons de croiser deux variables qualitatif, avec le sexe et l'état civil.

```
## Margins computed over dimensions
## in the following order:
## 1: Sexe
## 2: EtatCivil
```

Table 33: Tableau de contingence entre le sexe et l'état civil

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)	sum
Femme	7	3	38	5	53
Homme	17	10	86	2	115
sum	24	13	124	7	168

On commence par croiser les effectifs de nos deux variables dans le tableau 33. On apprend par exemple que sur les 168 hommes de nos données, 89 sont mariés.

```
## Margins computed over dimensions
## in the following order:
## 1: Sexe
## 2: EtatCivil
```

Table 34: Tableau des fréquences croisées entre le sexe et l'état civil (%)

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)	sum
Femme	4.17	1.79	22.62	2.98	31.55
Homme	10.12	5.95	51.19	1.19	68.45
sum	14.29	7.74	73.81	4.17	100.00

Le tableau 34 nous donnes les pourcentages du croisement entre nos deux variables. On remarque que 73.84% de nos individus sont mariés dont 22.62% sont des femmes.

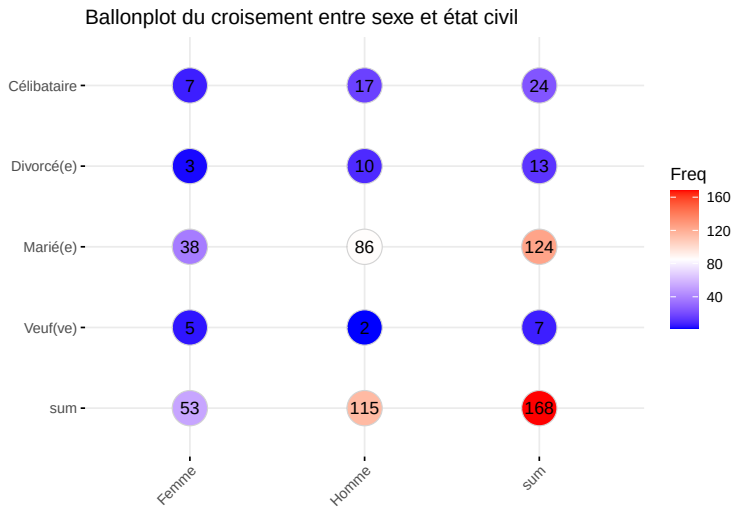

Figure 7: Ballonplot du croisement entre sexe et état civil

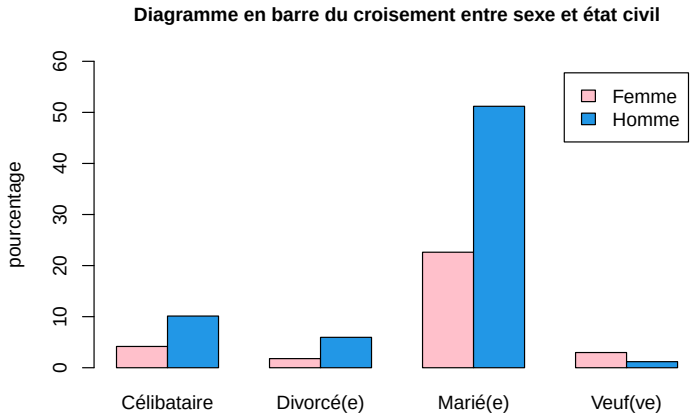

Figure 8: Diagramme en barre du croisement entre sexe et état civil

Nous pouvons rendre graphique les résultats de nos tableaux de contingence, comme le font les figures 7 et 8.

Table 35: Distribution conditionnelle du sexe sachant l'état civil (%)

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)	sum
Femme	29.17	23.08	30.65	71.43	31.55
Homme	70.83	76.92	69.35	28.57	68.45
sum	100.00	100.00	100.00	100.00	100.00

Le tableau 35 nous apprend que parmi nos individus veufs, 71.3% sont des femmes.

Table 36: Distribution conditionnelle de l'état civil sachant le sexe (%)

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)	sum
Femme	13.21	5.66	71.70	9.43	100
Homme	14.78	8.70	74.78	1.74	100
sum	14.29	7.74	73.81	4.17	100

Le tableau 36 nous dit que parmi nos individus femmes, 13.21% sont célibataires.

```
##
##  Pearson's Chi-squared test
##
## data:  TDC_E
## X-squared = 5.6972, df = 8, p-value = 0.6811
```

Quand on réalise le test pour savoir s'il y a indépendance entre nos deux variables, on trouve une pvaleur plutôt grande, ce qui ne nous permet pas de rejet H0, il y a indépendance entre le sexe et l'état civil.

Table 37: Tableau des effectifs théoriques

	Célibataire	Divorcé(e)	Marié(e)	Veuf(ve)	sum
Femme	7.57	4.1	39.12	2.21	53
Homme	16.43	8.9	84.88	4.79	115
sum	24.00	13.0	124.00	7.00	168

Grâce au tableau 37 on peut obtenir la statistique de notre test. On trouve 5.69 ce résultat est visible dans les sorties de notre de test. Le quantile de la loi de χ^2 est de 7.81, plus grand que notre stat de test. On ne rejet donc pas l'hypothèse d'indépendance.

Croisement quantitatif vs qualitatif : Stress vs EtatCivil

Maintenant croisons l'état civil avec une autre variable qui est le stress.

Table 38: Statistique élémentaire de la variable stress

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.7	15.185	18.19	18.2044	21.115	31.84

Regardons les statistiques élémentaires de cette variable. Avec le tableau 8 on remarque par exemple que le stress moyen est de 18.20.

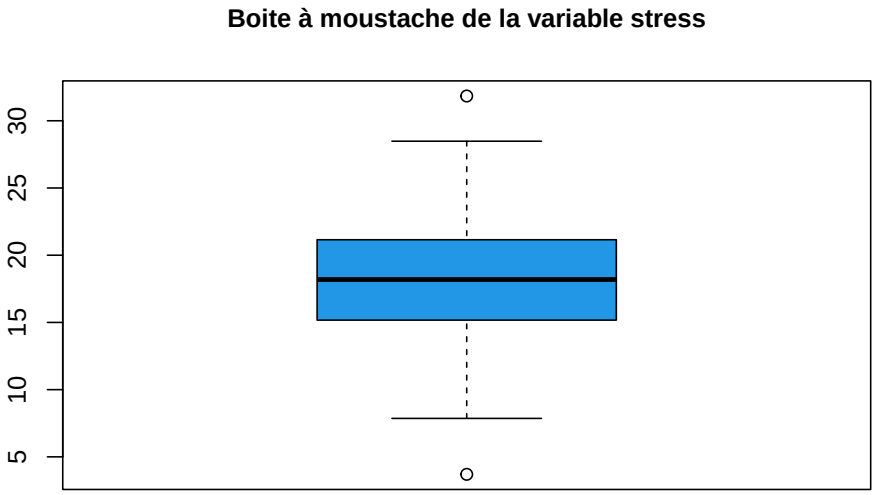

Figure 9: Boite à moustache de la variable stress

Illustrons ces statistiques. Sur la figure 9 on voit que cette variable est bien distribuée autour de la médiane. On voit également qu'il y a 2 valeurs aberrantes.

```
## Margins computed over dimensions
## in the following order:
## 1: EtatCivil
## 2: decoup_stress
```

Table 39: Croisement entre l'état civil et le stress

	(3.67,9.33]	(9.33,15]	(15,20.6]	(20.6,26.2]	(26.2,31.9]	sum
Célibataire	2	4	15	3	0	24
Divorcé(e)	1	3	7	1	1	13
Marié(e)	2	23	58	34	7	124
Veuf(ve)	1	4	2	0	0	7
sum	6	34	82	38	8	168

Pour analyser le croisement entre ces deux variables il faut au préalable découper en classes la variable stess. Nous la découpons en 5 classes de même amplitude. Une fois réalisé nous pouvons faire le tableau 39 du croisement des effectifs. On apprend par exemple que 58 individus mariés ont un stress entre 15 et 20.6.

```
## Margins computed over dimensions
## in the following order:
## 1: EtatCivil
## 2: decoup_stress
```

Table 40: Fréquence croisées en %

	(3.67,9.33]	(9.33,15]	(15,20.6]	(20.6,26.2]	(26.2,31.9]	sum
Célibataire	1.190	2.381	8.929	1.786	0.000	14.286
Divorcé(e)	0.595	1.786	4.167	0.595	0.595	7.738
Marié(e)	1.190	13.690	34.524	20.238	4.167	73.809
Veuf(ve)	0.595	2.381	1.190	0.000	0.000	4.166
sum	3.570	20.238	48.810	22.619	4.762	99.999

Le tableau 40 nous apprend que 2.38% des célibataires ont un stress entre 9.33 et 15.

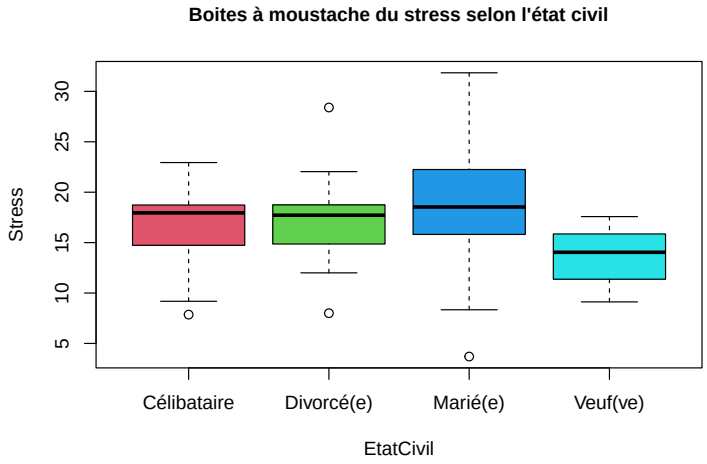

Figure 10: Boites à moustache du stress selon l'état civil

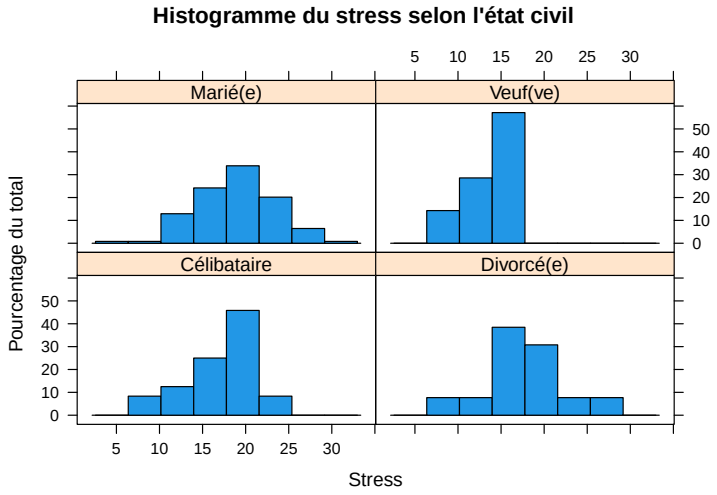

Figure 11: Histograme du stress selon l'état civil

On remarque avec la figure 10 et 11 que la distribution du stress entre les célibataires et les divorcés est assez similaire. Pour les mariés on retrouve une distribution étendue qui va prendre des valeurs plus grandes. Alors que pour les veufs on retrouve un étendu plus faible, les valeurs du stress pour les veufs sont plus faibles.

```
## EtatCivil: Célibataire
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.86   14.91   17.95   16.76   18.66   22.94
## ------------------------------------------------------------
## EtatCivil: Divorcé(e)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   14.86   17.72   17.17   18.74   28.40
## ------------------------------------------------------------
## EtatCivil: Marié(e)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.70   15.82   18.53   18.85   22.25   31.84
## ------------------------------------------------------------
## EtatCivil: Veuf(ve)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.12   11.37   14.04   13.60   15.86   17.58
```

On peut regarder les statistiques élémentaires du stress selon l'état civil. On voit par la médiane pour les veufs est plus faible que pour les autres groupes. On retrouve les mêmes résultats que la figure 10.

On peut calculer le rapport de corrélation entre nos deux variables. On trouve 0.075. Cela nous dit que 7.5% de la variabilité du stress est expliquée par l'état civil.

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## EtatCivil     3    265   88.34   4.428 0.00506 **
## Residuals   164   3272   19.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On réalise un test pour savoir s'il y a une différence du stress entre les états civils. On trouve une pvaleur inférieure à 0.05, on peut rejeter l'hypothèse nul et dire qu'il y a une différence entre la valeur du stress selon les différents états civils.

Croisement quantitatif vs quantitatif : Age vs Satisfaction

On va maintenant voir le croisement entre les variables âge et satisfaction.

Table 41: Statistiques élémentaires de la satisfaction

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.85	13.8375	19.17	20.43101	28.3125	38.45

Table 42: Statistiques élémentaires de l'âge

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
25	37	41	41.9881	49.25	57

Les tableaux 41 et 42 nous donnes les statistiques élémentaires sur nos variables. On remarque par exemple que la satisfaction moyenne et de 20.43 et celle de l'âge est de 41.99.

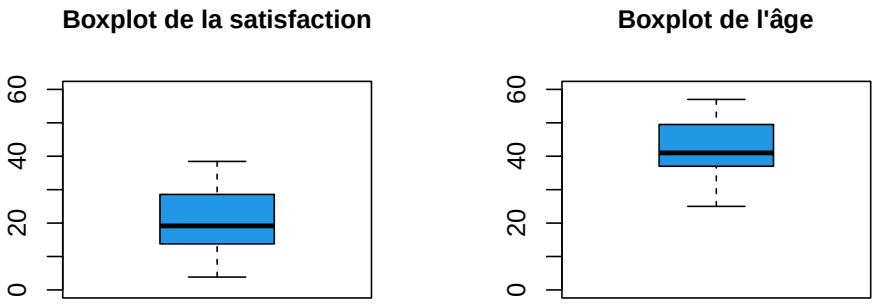

Figure 12: Boxplot des variable satisfaction et age

On peut rend visuel nos résultats comme dans la figure 12. On remarque que, pour les deux variables, il n'y a pas de valeur aberrante.

```
## Margins computed over dimensions
## in the following order:
## 1: decoup_Satisfaction
## 2: decoup_age
```

Table 43: Tableau de contingence entre la satisfaction et l'age

	(25,33]	(33,41]	(41,49]	(49,57]	sum
(3.82,10.8]	26	1	0	0	27
(10.8,17.7]	2	38	3	0	43
(17.7,24.6]	0	20	22	0	42
(24.6,31.5]	0	1	11	18	30
(31.5,38.5]	1	0	1	24	26
sum	29	60	37	42	168

Nous avons deux variables quantitatives. Il faut donc créer des classes pour chacune d'entre elles, afin de pouvoir les croiser. Quand on croise les effectifs, on obtient la distribution résumée dans le tableau 43. On apprend par exemple que 38 de nos individus qui ont entre 33 et 41 ans ont une satisfaction comprise ente 10.8 et 17.7. Si on est bien attentif on remarque que quand l'âge augmente, la satisfaction augmente également. Regardons cela dans un graphique.

```
## Margins computed over dimensions
## in the following order:
## 1: decoup_Satisfaction
## 2: decoup_age
```

Table 44: Fréquence croisées entre la satisfaction et l'age (%)

	(25,33]	(33,41]	(41,49]	(49,57]	sum
(3.82,10.8]	15.5	0.6	0.0	0.0	16.1
(10.8,17.7]	1.2	22.6	1.8	0.0	25.6
(17.7,24.6]	0.0	11.9	13.1	0.0	25.0
(24.6,31.5]	0.0	0.6	6.5	10.7	17.8
(31.5,38.5]	0.6	0.0	0.6	14.3	15.5
sum	17.3	35.7	22.0	25.0	100.0

Avant cela, regardons d'autres façon d'illustrée le croisement entre nos deux variables. Avec le tableau 44 on retrouve le pourcentage de chaque croisement.

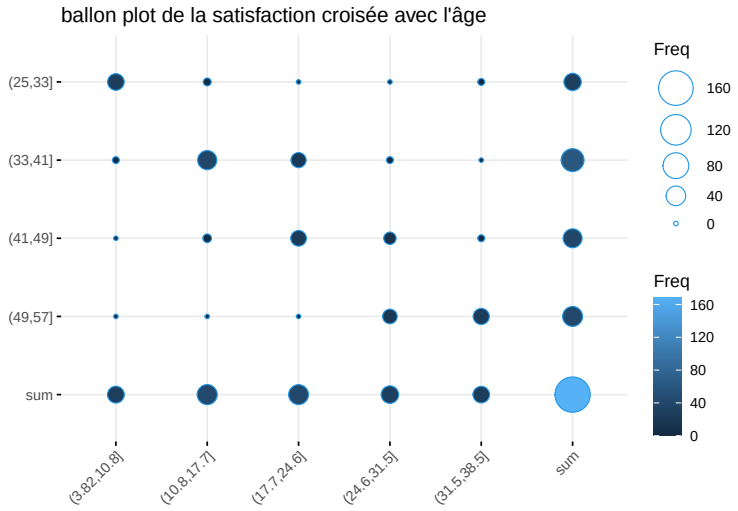

Figure 13: ballon plot de la satisfaction croisée avec l'âge

Ou encore avec la figure 13 qui donne un cercle plus ou moins gros selon l'effectif du croisement.

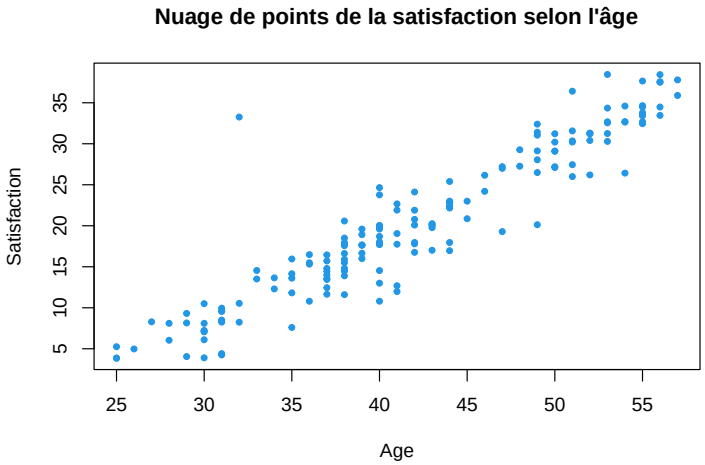

Figure 14: Nuage de points de la satisfaction selon l'age

Graphiquement on se rend compte directement de la liaison entre nos deux variables. La figure 14 nous montre une relation qui est croissante est positive. On voit un point qui ne se comporte pas comme les autres, un intrus.

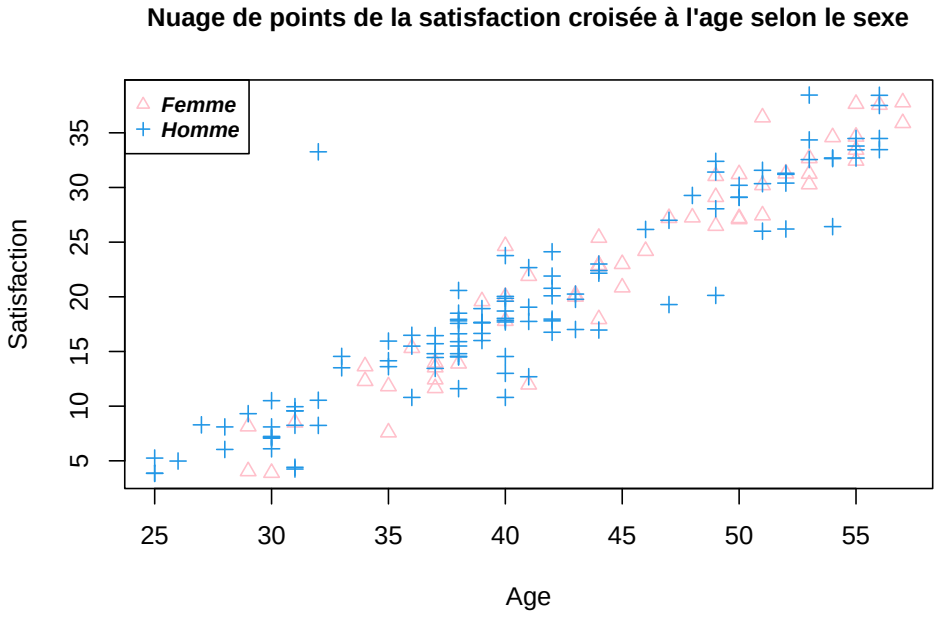

Figure 15: Nuage de points de la satisfaction croisée à l'âge selon le sexe

On peut ajouter l'information sur le sexe comme sur la figure 15. Le sexe ne semble pas avoir d'influence sur notre croisement de la satisfaction et de l'âge.

Table 45: Matrice des corrélations

	Anciennete	Salaire	Satisfaction	Stress	EstimeSoi	Age
Anciennete	1.000	0.055	1.000	-0.620	0.995	0.932
Salaire	0.055	1.000	0.074	0.006	0.049	0.056
Satisfaction	1.000	0.074	1.000	-0.612	0.995	0.938
Stress	-0.620	0.006	-0.612	1.000	-0.666	-0.420
EstimeSoi	0.995	0.049	0.995	-0.666	1.000	0.935
Age	0.932	0.056	0.938	-0.420	0.935	1.000

Le tableau 45 nous donne les liens entre chaque variable. Plus le chiffre est proche de 1 plus les liens et fort positivement, plus il est proche de -1 plus le lien et fort mais négativement. Vers 0 la relation est faible. Pour la satisfaction et l'âge on voit un coefficient très proche de 1, ce qui confirme les résultats vus précédemment.

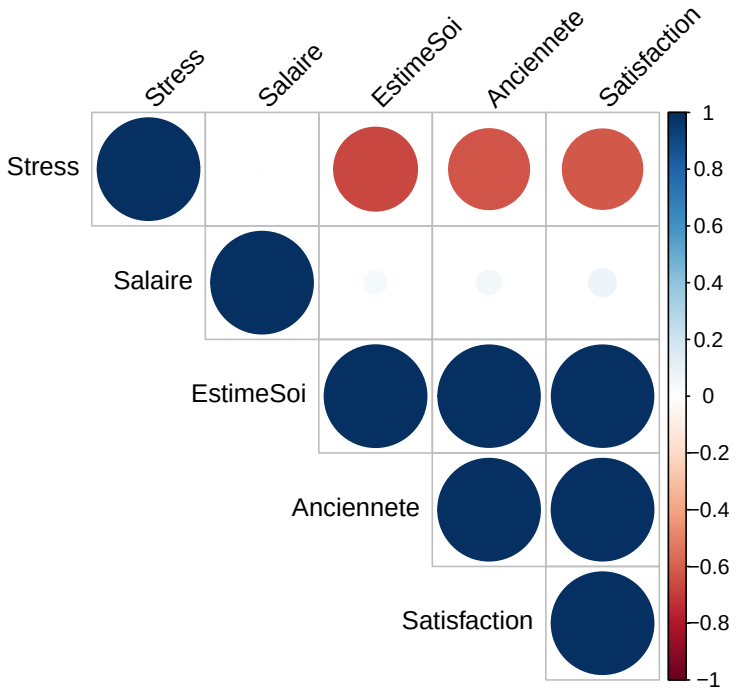

Figure 16: corrélogramme

On peut illustrer ce tableau comme le montre la figure 16. Plus le cercle est bleu plus la corrélation est forte, comme par exemple avec l'ancienneté et la satisfaction.

Chapitre 3 : Analyse en Composantes Principales

Exercice 23

L'objectif va être de retrouver les données trouvées en cours sur un jeu de données. Les informations présentent dans le cours sont sur ces données : la matrice de corrélation, le vecteur propre, les valeurs propres, ainsi que les coordonnées pour les variables et individus.

Table 46: Données

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

Le tableau 46 nous montre les données utilisées, il y a 2 variables, Z1 et Z2, pour 5 individus.

Table 47: Données centrée réduite

X1	X2
-1.004	-1.545
-0.645	0.601
-0.287	-0.258
0.072	-0.258
1.864	1.459

Pour étudier ces données il faut d'abord les centrer et réduire. Nous obtenons donc les données du tableau 47 avec X1 et x2 comme nouvelles variables. Dans le cours les écarts types utilisés sont faux. Ici on utilisera les écarts types du cours afin d'avoir les mêmes résultats que dans le cours.

Table 48: Matrice des corrélations

	X1	X2
X1	1.000	0.788
X2	0.788	1.000

La première matrice que nous avons est la matrice des corrélations. On trouve avec le tableau 48 un rapport de corrélation de 0.788 comme dans le cours. Il y a donc une légère corrélation positive entre nos deux variables.

Table 49: Vecteur Propre

0.707	-0.707
0.707	0.707

On calcule ensuite le vecteur propre, qui constitue le tableau 49. On trouve une différence par rapport aux résultats du cours. En effet le signe moins n'est pas sur la même valeur dans le cours.

Le premier facteur associé à la valeur propre 1.787, et 0.212 pour le deuxième. On va donc retrouver les mêmes valeurs de pourcentages de variance expliquées, 89.4% pour l'axe 1 et 10,6% pour l'axe 2. On conservera ces deux axes.

Table 50: Coordonnées des variables

	Cord_Axe1	Cord_Axe2
X1	0.946	-0.326
X2	0.946	0.326

On peut calculer les coordonnées de nos deux variables. Nous trouvons avec le tableau 50 que les coordonnées de X1 seront de 0.944 sur l'axe 1 et -0.326. Et pour X2, 0.947 sur l'axe 1 et 0.325 pour l'axe 2. Par rapport aux résultats du cours on retrouve le même problème vis-à-vis du signe moins.

Table 51: Coordonnées des individus

X1	X2
-1.802	-0.383
-0.031	0.881
-0.385	0.021
-0.132	-0.233
2.350	-0.286

Il nous reste plus que les coordonnées des individus. Ont les retrouvent dans le tableau 51. Les individus 1 et 5 contribuent fortement à l'axe 1, tandis que l'individu 2 contribue le plus à l'axe 2.

Nous allons maintenant comparer 2 fonctions qui réalisent des ACP sous R. Il s'agit de princomp et prcomp. On conservera les données précedentes.

Table 52: Valeurs propres avec princomp

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.788	89.401	89.401
Dim.2	0.212	10.599	100.000

Table 53: Valeurs propres avec prcomp

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.788	89.401	89.401
Dim.2	0.212	10.599	100.000

On commene par les valeurs propres. On voit les sorties des deux fonctions dans les tableaux 52 et 53, et on remarque que les sorties sont exactement les mêmes.

Table 54: coordonnées des variables avec princomp

	Dim.1	Dim.2
Z1	0.946	0.326
Z2	0.946	-0.326

Table 55: coordonnées des variables avec prcomp

	Dim.1	Dim.2
Z1	0.946	0.326
Z2	0.946	-0.326

Pour les coordonnées des variables on voit dans les tableaux 54 et 55, les sorties des deux fonctions sont les mêmes. A noter qu'avec ces deux fonctions le signe moins est comme dans le cours. On va voir que cela est différent avec la fonction PCA.

Table 56: Cos2 des variables avec princomp

	Dim.1	Dim.2
Z1	0.894	0.106
Z2	0.894	0.106

Table 57: Cos2 des variables avec prcomp

	Dim.1	Dim.2
Z1	0.894	0.106
Z2	0.894	0.106

Pour les qualités de représentations des variables (cos2) des variables on voit dans les tableaux 10 et 11, les sorties des deux fonctions sont les mêmes.

Table 58: Contributions des variables avec princomp

	Dim.1	Dim.2
Z1	50	50
Z2	50	50

Table 59: Contributions des variables avec prcomp

	Dim.1	Dim.2
Z1	50	50
Z2	50	50

Et pour finir les contributions des variables sont aussi la mêmes entre les deux fonctions.

Table 60: Coordonnées des individus avec princomp

Dim.1	Dim.2
-1.802	0.381
-0.032	-0.881
-0.385	-0.021
-0.131	0.233
2.351	0.289

Table 61: Coordonnées des individus avec prcomp

Dim.1	Dim.2
-1.612	0.340
-0.029	-0.788
-0.344	-0.019
-0.117	0.208
2.103	0.259

Pour les individus maintenant. C'est là que l'on retrouve des différences. Déjà pour les coordonnées on trouve des dissimilarités on le voit dans les tableaux 60 et 61.

Table 62: Cos2 des individus avec princomp

Dim.1	Dim.2
0.957	0.043
0.001	0.999
0.997	0.003
0.241	0.759
0.985	0.015

Table 63: Cos2 des individus avec prcomp

Dim.1	Dim.2
0.957	0.043
0.001	0.999
0.997	0.003
0.241	0.759
0.985	0.015

Cependant les qualités de représentation des individus (cos2) sont les mêmes d'une fonction à une autre.

Table 64: Contributions des individus avec princomp

Dim.1	Dim.2
36.328	13.672
0.012	73.288
1.658	0.042
0.192	5.108
61.810	7.890

Table 65: Contributions des individus avec prcomp

Dim.1	Dim.2
29.062	10.938
0.009	58.631
1.326	0.034
0.154	4.086
49.448	6.312

Et pour les contributions, on remarque dans les tableaux ci-dessus que les valeurs sont différentes.

Les plus grosses différences entre princomp et prcomp ce font au niveau des individus et non pour les variables. Regardons maintenant les résultats avec la fonction PCA.

Table 66: Valeurs propres avec PCA

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.788	89.401	89.401
Dim.2	0.212	10.599	100.000

Pour les valeurs propres, on retrouve encore les résultats précédents comme le montre le tableau 66.

Table 67: coordonnées des variables avec PCA

	Dim.1	Dim.2
Z1	0.946	-0.326
Z2	0.946	0.326

Pour les coordonnées des variables on voit dans le tableau 67 que le signe moins est placé comme pour l'ACP à la main. Plus exactement ce signe change de position si on utilise les données centrées réduites, hors ceci est fait automatiquement dans cette fonction.

Table 68: Cos2 des variables avec PCA

	Dim.1	Dim.2
Z1	0.894	0.106
Z2	0.894	0.106

Table 69: Contributions des variables avec PCA

	Dim.1	Dim.2
Z1	50	50
Z2	50	50

Pour les qualités de représentation des variables et les contributions, pas de problème on retrouve les mêmes résultats avec les 3 fonctions.

Table 70: Coordonnées des individus avec PCA

Dim.1	Dim.2
-1.802	-0.381
-0.032	0.881
-0.385	0.021
-0.131	-0.233
2.351	-0.289

Maintenant regardons les résultats de PCA avec les individus. Pour les coordonnées la fonction PCA s'accorde avec la fonction princomp.

Table 71: Cos2 des individus avec PCA

Dim.1	Dim.2
0.957	0.043
0.001	0.999
0.997	0.003
0.241	0.759
0.985	0.015

Pour le cos2, la fonction PCA s'accorde avec toutes les autres méthodes.

Table 72: Contributions des individus avec PCA

Dim.1	Dim.2
36.328	13.672
0.012	73.288
1.658	0.042
0.192	5.108
61.810	7.890

Et pour finir le tableau ci-dessus nous montre que les contributions de la fonction PCA sont les mêmes qu'avec la fonction princomp.

On en déduit que les fonctions PCA et princomp réalise l'ACP de la même façon.

Exercice 24

Dans cette partie on traitera des données des stations de ski en Savoie. On dispose, pour 32 stations, des variables suivantes (données 1998).

— prixforf : prix du forfait 1 semaine (Euros)

— altmin : altitude minimum de la station (m)

— altmax : altitude maximum de la station (m)

— pistes : nombre de pistes de ski alpin

— kmfond : nombre de kilomètres de pistes de ski de fond remontee : nombre de remontées mécaniques

Table 73: Extrait des données

prixforf	altmin	altmax	pistes	kmfond
76	900	2000	45	50
160	800	3226	117	30
85	750	2300	30	47
71	500	2750	21	10
54	1710	2200	4	80
79	1850	3000	16	0

Voici un extrait des données que nous traiterons.

Valeurs propres :

Les valeurs propres mesurent la quantité de variance expliquée par chaque axe principal. Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération.

Table 74: Valeurs propres

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.393	47.850	47.850
Dim.2	1.214	24.282	72.132
Dim.3	0.850	16.998	89.130
Dim.4	0.429	8.589	97.719
Dim.5	0.114	2.281	100.000

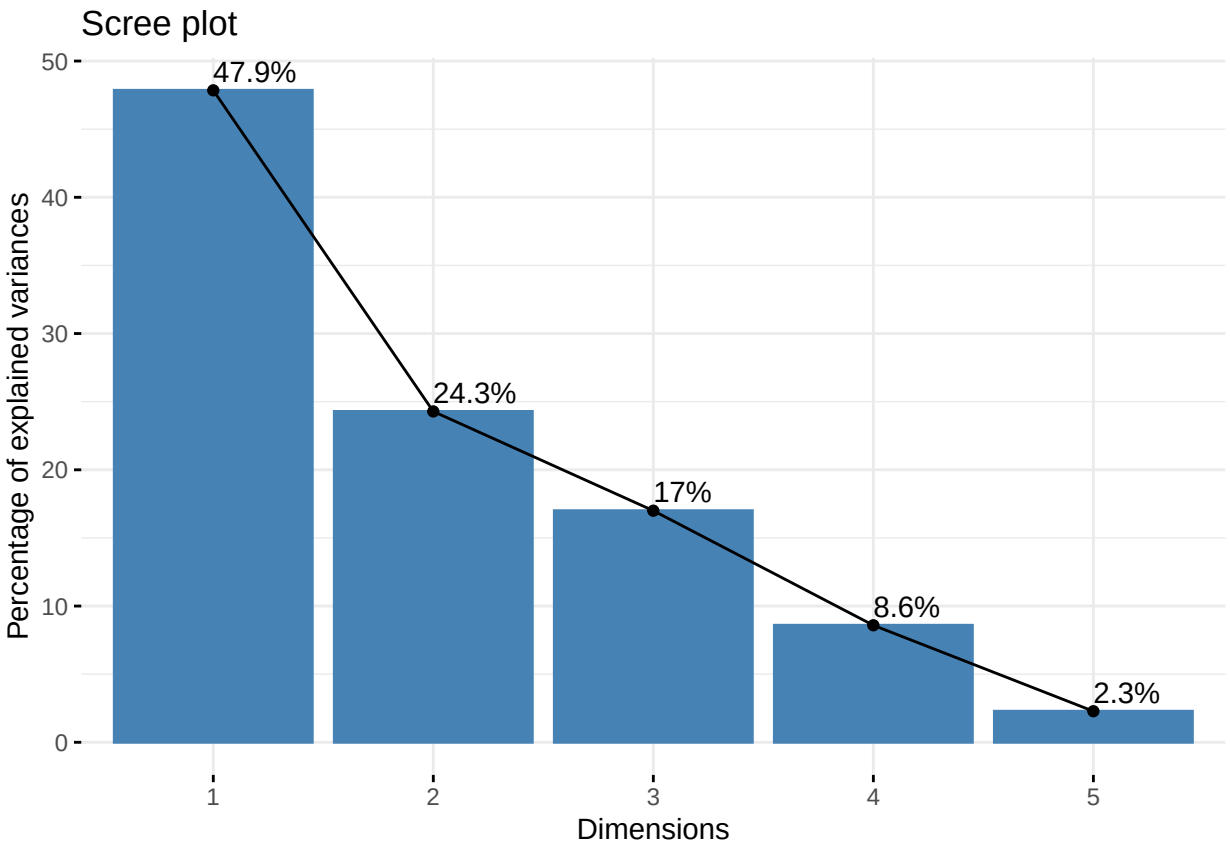

Figure 17: Visualisation des valeurs propres

On voit avec le tableau 74 et la figure 17 qu'avec 3 axes on obtient une variance expliquée de presque 90%, ce qui est suffisant. On conservera donc les 3 premières dimensions.

Variables :

Table 75: coordonnées des variables

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
prixforf	0.938	0.001	-0.102	0.230	-0.238
altmin	-0.001	0.830	0.509	0.228	0.029
altmax	0.736	0.440	-0.052	-0.512	0.010
pistes	0.926	-0.147	-0.126	0.219	0.237
kmfond	0.335	-0.558	0.750	-0.122	-0.012

Le tableau suivant montre les valeurs des coordonnées afin de créer un cercle des corrélations.

Table 76: Qualité de représentation des variables

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
prixforf	0.880	0.000	0.010	0.053	0.057
altmin	0.000	0.688	0.259	0.052	0.001
altmax	0.542	0.193	0.003	0.262	0.000
pistes	0.858	0.022	0.016	0.048	0.056
kmfond	0.112	0.311	0.562	0.015	0.000

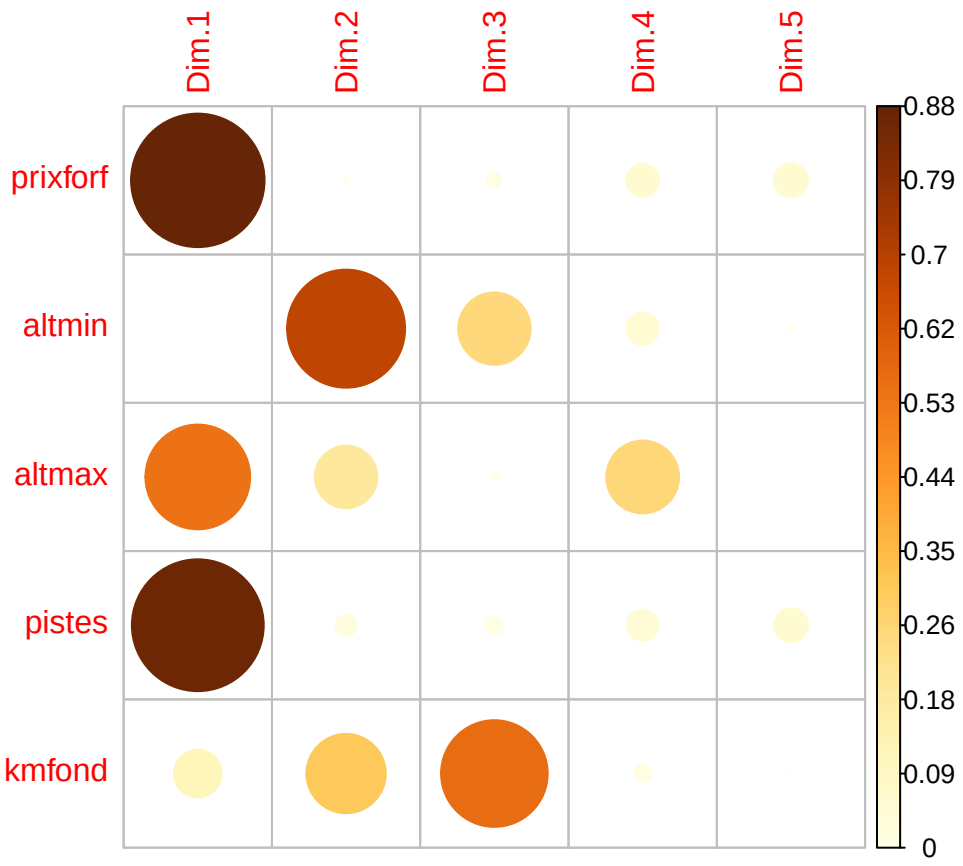

Figure 18: Corrélogramme des variables pour chaque dimension

Ensuite on regarde les qualités de représentations des variables pour chaque dimension à l'aide du graphique et tableau ci-dessus. On voit par exemple que les variables prixforf et piste sont très bien représentées sur l'axe 1.

Table 77: Contributions des variables

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
prixforf	36.790	0.000	1.213	12.300	49.698
altmin	0.000	56.682	30.484	12.081	0.753
altmax	22.646	15.923	0.318	61.023	0.090
pistes	35.875	1.792	1.874	11.117	49.343
kmfond	4.689	25.604	66.112	3.479	0.116

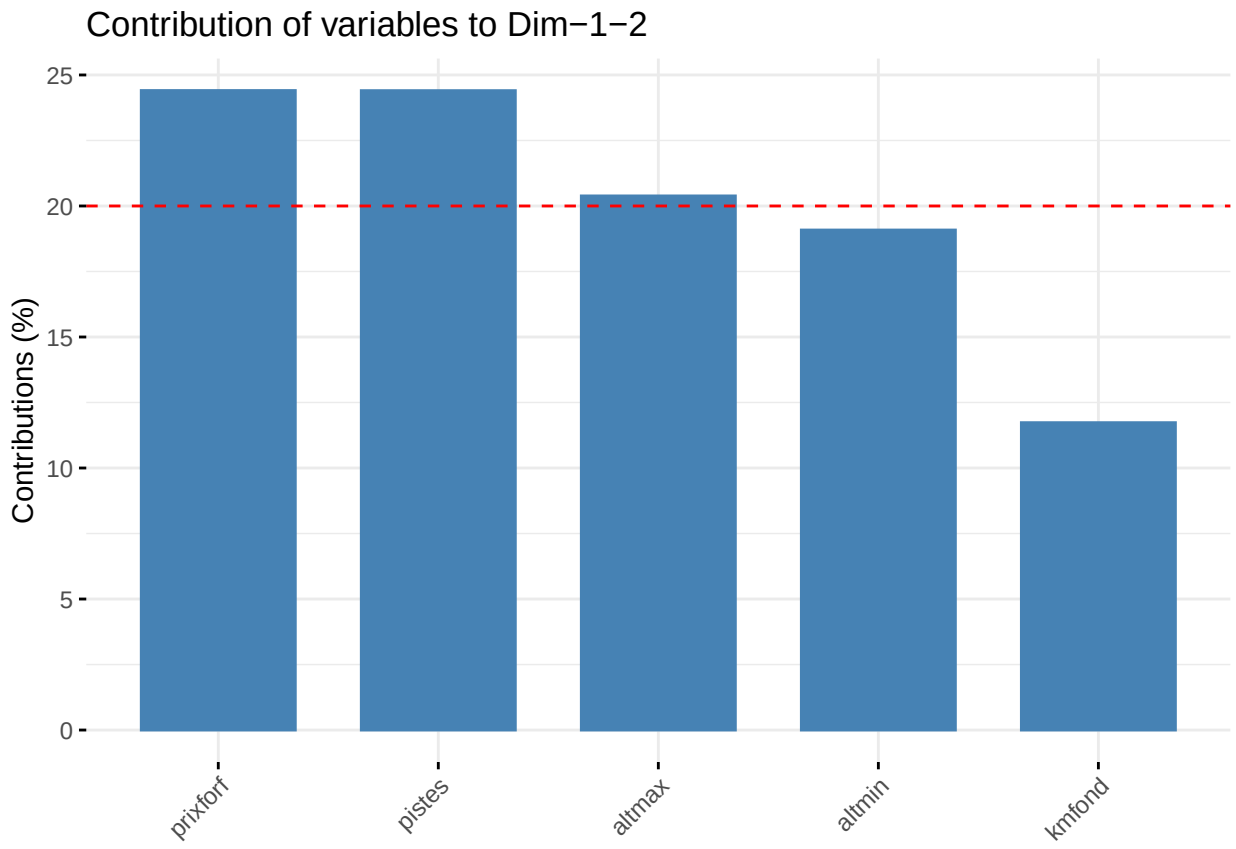

Figure 19: Diagramme en barre des contributions des variables

On s'intéresse maintenant à la contribution des variables. On voit avec le tableau 77 les contributions de chaque variable pour chaque dimension. Par exemple la variable prixforf contribue à 36.7% de l'axe 1. Le graphique nous montre ces contributions pour le premier plan. La ligne rouge indique la contribution moyenne attendue. Toutes les variables qui dépassent cette ligne sont considérées comme importantes pour contribuer au premier plan.

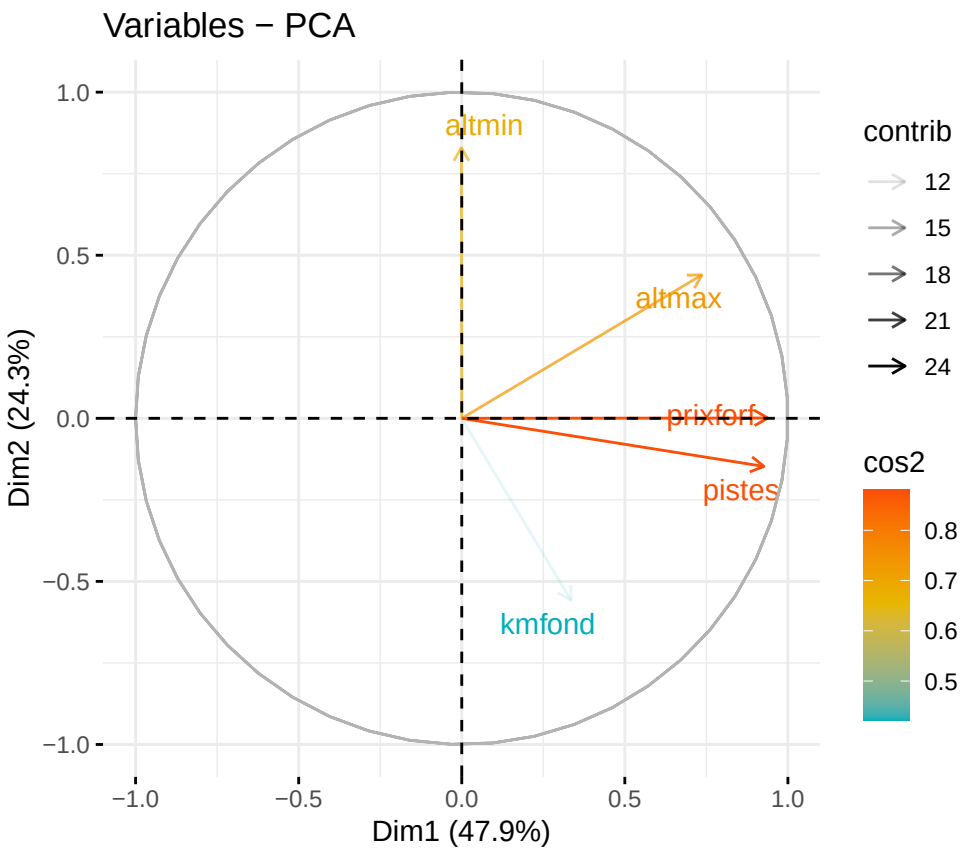

Figure 20: Cercle de corrélation des variables

On peut représenter l'ensemble de ces résultats dans le cercle des corrélations, ici on s'interesse au premier plan. Pour la qualité de représentation (cos2) on a une variation de couleur selon son importance sur le premier plan. On voit que les variables prixfof et piste sont celles qui sont le mieux représenter dans ce plan. la transparence des flèches indique la contribution. Par exemple la variable kmfond n'a pas une grande contribution sur le premier plan. Les flèches des variables prixfof et piste sont très proches ce qui indique un lien fort positif entre ces deux variables. Tandis que atmin et kmfond son pratiquement opposée ce qui signifie un lien fort négatif.

Individus :

Table 78: Extrait des coordonnées des individus

Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
-0.967	-1.924	0.276	-0.066	0.358
2.908	-0.969	-1.327	-0.383	0.073
-0.790	-1.893	-0.089	-0.736	-0.164
-1.118	-1.234	-1.800	-1.671	-0.024
-1.619	-0.408	2.926	-0.421	0.180
-0.905	2.310	0.075	-0.529	0.085

On s'intéresse maintenant aux individus. On commence par montrer leurs coordonnées qu'ils auront sur chaque dimension sur le tableau ci-dessus. Ici il n'y a pas l'ensemble des individus représentés.

Table 79: Extrait des cos2 des individus

Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
0.193	0.764	0.016	0.001	0.026
0.748	0.083	0.156	0.013	0.000
0.130	0.749	0.002	0.113	0.006
0.142	0.173	0.368	0.317	0.000
0.227	0.014	0.741	0.015	0.003
0.127	0.828	0.001	0.043	0.001

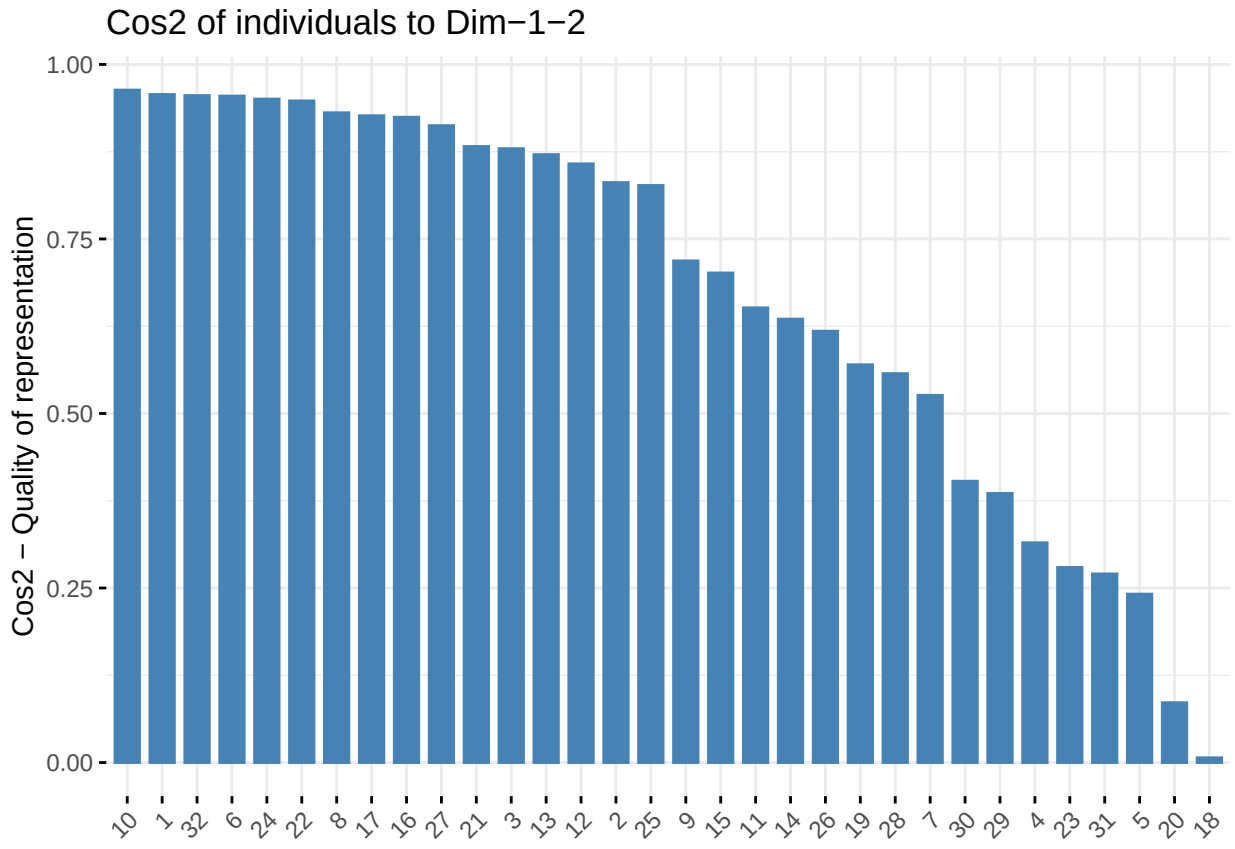

Figure 21: Diagramme en barre des cos2 des individus

On regarde maintenant les qualités de représentation pour les différentes dimensions, qu'on retrouve dans le tableau. Le graphique nous illustre ces valeurs pour le premier plan, on voit que c'est l'individu 10 qui est le mieux représenter sur ce plan, suivis du 1 et du 32. Nous allons les retrouver après dans le graphique.

Table 80: Extrait des contributions des individus

Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1.221	9.527	0.280	0.032	3.512
11.044	2.416	6.476	1.070	0.145
0.815	9.220	0.029	3.946	0.739
1.631	3.919	11.918	20.316	0.016
3.424	0.429	31.469	1.287	0.892
1.069	13.738	0.021	2.035	0.197

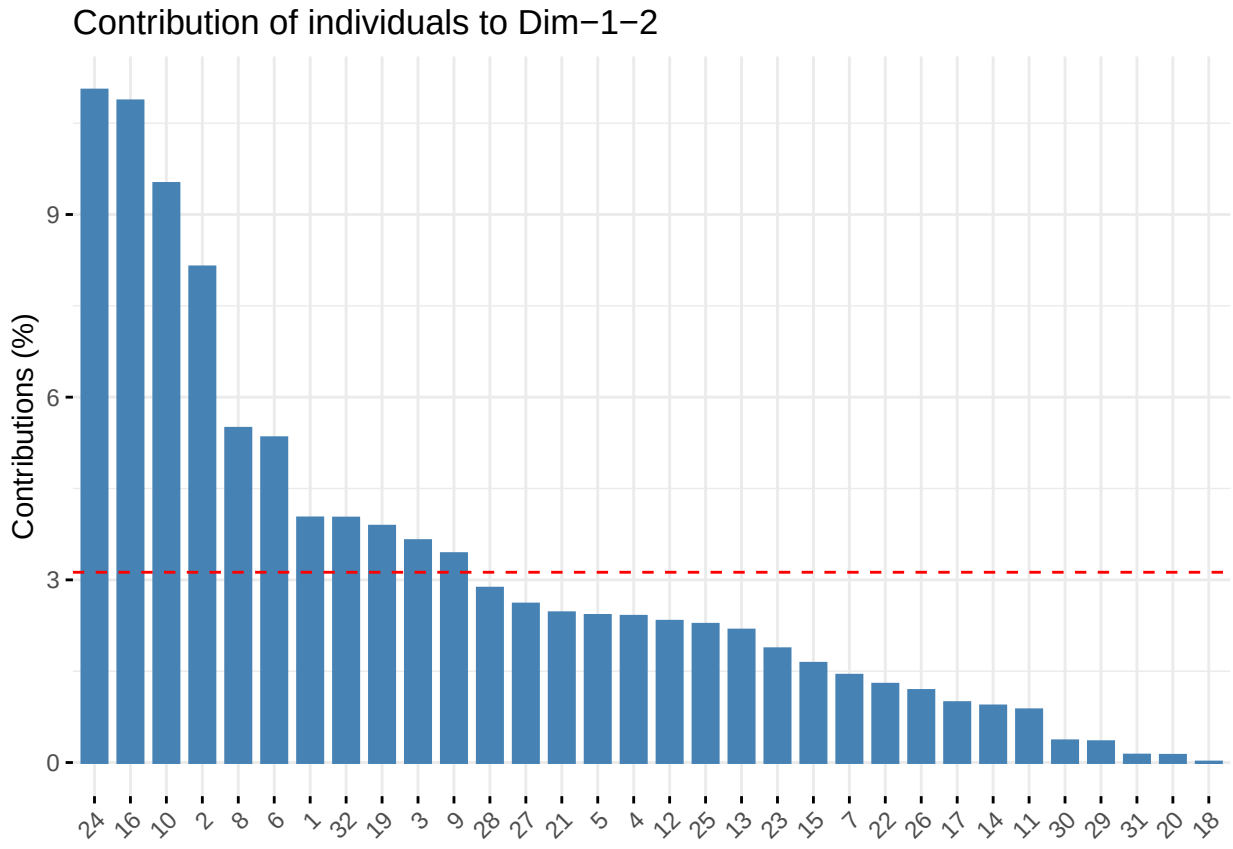

Figure 22: Diagramme en barre de la contributions des individus

On s'intéresse maintenant à la contribution des individus. On voit dans le tableau les contributions des individus sur les différents axes. Le graphique montre les individus qui contribuent le plus au premier plan. L'individu 24 est celui qui a la plus forte contribution sur ce plan. Tous les individus au-dessus de la ligne pointillée rouge peuvent être considérés comme importants pour contribuer au premier plan.

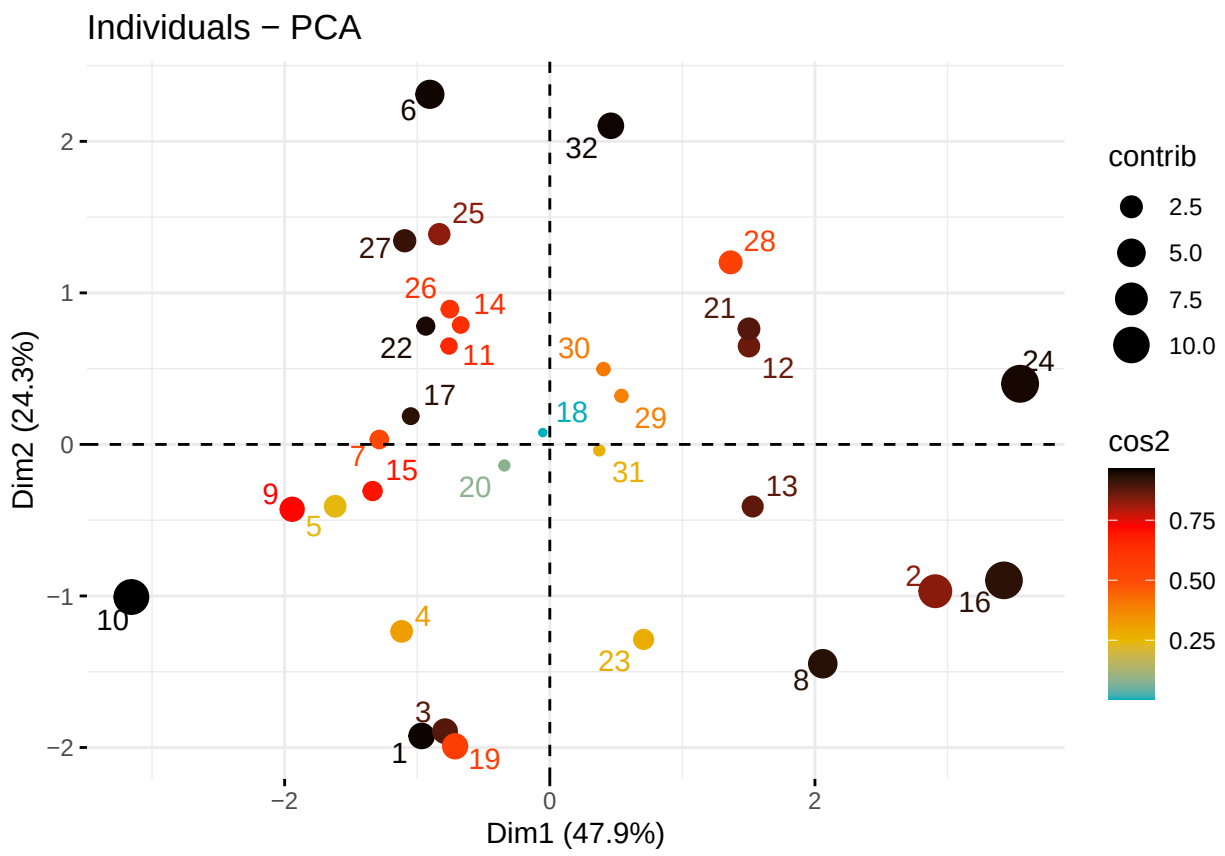

Figure 23: Nuage de points des individus

On peut maintenant tracer le nuage de points des individus sur le premier plan. Chaque point à une épaisseur proportionnelle à sa contribution. On retrouve les individus 24, 10 et 16 qui sont les plus gros points et qui contribuent le plus. De plus nous pouvons voir une couleur plus chaude quand la qualitée de représentation (cos2) est élevée, et inversement. Les individus 10,1,24,6,22,8, et 32 sont de couleur noire qui correspond à une très bonne qualité de représentation sur ce plan.

Variables & individus :

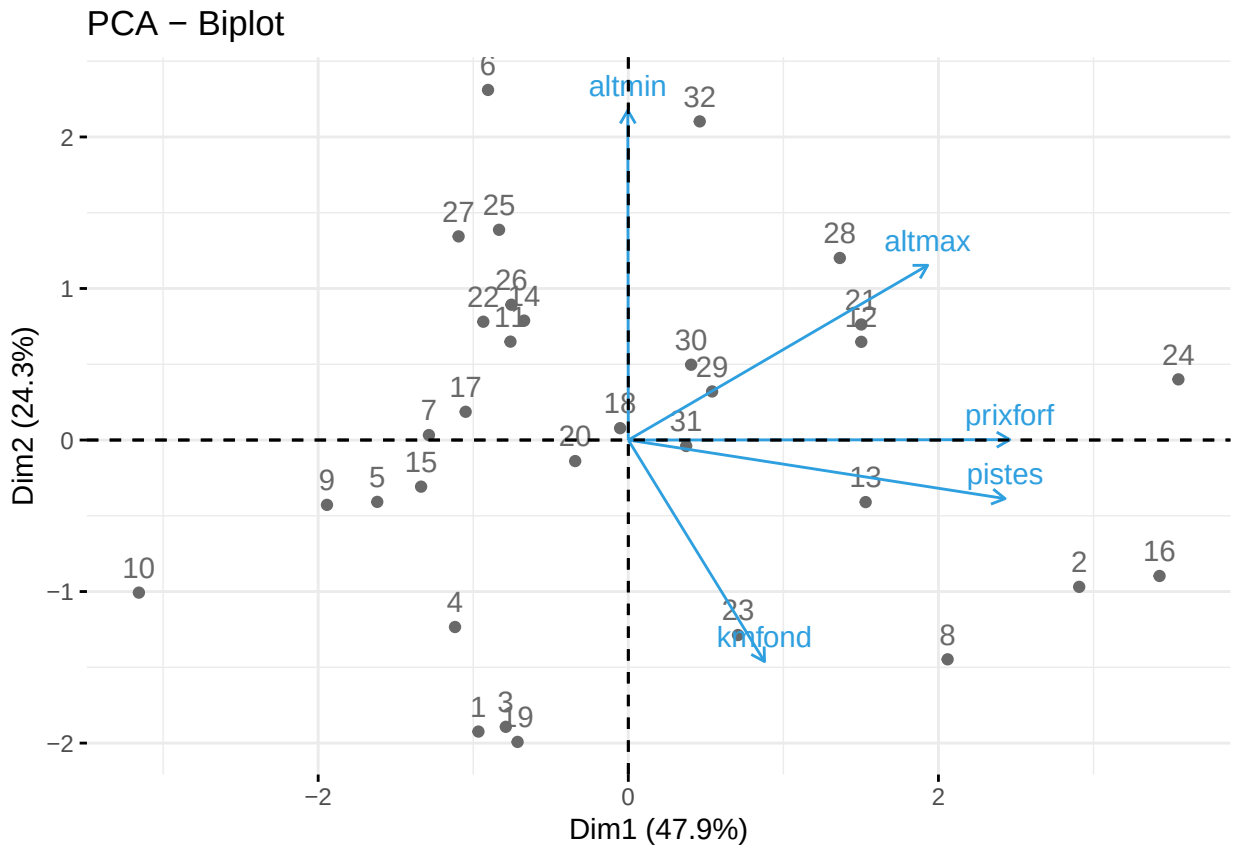

Figure 24: Bitplot

Pour finir on va analyser les individus et les variables ensembles sur le premier plan, en supposant les deux graphiques précédents. Le graphique nous apprend quels sont les individus qui ont une forte valeur selon les variables. Pour la variable altmax par exemple, les individus 29, 12, 21, sont ceux qui ont une forte valeur pour cette variable. Et inversement, on peut savoir quels individus ont une faible valeur pour une variable. Les individus 27, 25, 26 par exemple sont opposés à la flèche de la variable kmfond, ceux qui indiquent une faible valeur de ces individus pour cette variable.

Chapitre 4 : Analyse Factorielle des Correspondances (AFC)

Exercice 31

Nous travaillons avec le jeu de données USArrests disponible dans R. Ces données contiennent des statistiques, en nombre d'arrestations pour 100 000 résidents pour agression, meurtre et viol dans chacun des 50 États américains en 1973. Le pourcentage de la population vivant dans des zones urbaines est également indiqué.

Murder : Nombre d'arrestations pour meurtre (pour 100 000 résidents)

Assault : Nombre d'arrestations pour agression (pour 100 000 résidents)

UrbanPop : Pourcentage de la population urbaine

Rape : Nombre Arrestations pour viols (pour 100 000 résidents)

L'objectif va être de comparer les sorties de l'ACP sur ces données avec 3 fonctions différentes, PCA, prcomp, et princomp.

Table 81: Extrait des Coordonnées avec PCA

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.986	-1.133	0.444	0.156
Alaska	1.950	-1.073	-2.040	-0.439
Arizona	1.763	0.746	-0.055	-0.835
Arkansas	-0.141	-1.120	-0.115	-0.183
California	2.524	1.543	-0.599	-0.342

Table 82: Extrait des Coordonnées avec prcomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	-0.976	1.122	-0.440	0.155
Alaska	-1.931	1.062	2.020	-0.434
Arizona	-1.745	-0.738	0.054	-0.826
Arkansas	0.140	1.109	0.113	-0.181
California	-2.499	-1.527	0.593	-0.339

Table 83: Extrait des Coordonnées avec princomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.986	1.133	0.444	0.156
Alaska	1.950	1.073	-2.040	-0.439
Arizona	1.763	-0.746	-0.055	-0.835
Arkansas	-0.141	1.120	-0.115	-0.183
California	2.524	-1.543	-0.599	-0.342

On remarque que les sorties des coordonnées sont quasiment les mêmes entre PCA et princomp, il y a que sur la dimension 2 que le signe est différent. La fonction prcomp elle prend des valeurs totalement diférentes.

Table 84: Extrait des Cos2 avec PCA

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.392	0.518	0.080	0.010
Alaska	0.409	0.124	0.447	0.021
Arizona	0.712	0.127	0.001	0.160
Arkansas	0.015	0.950	0.010	0.025
California	0.690	0.258	0.039	0.013

Table 85: Extrait des Cos2 avec prcomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.392	0.518	0.080	0.010
Alaska	0.409	0.124	0.447	0.021
Arizona	0.712	0.127	0.001	0.160
Arkansas	0.015	0.950	0.010	0.025
California	0.690	0.258	0.039	0.013

Table 86: Extrait des Cos2 avec princomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.392	0.518	0.080	0.010
Alaska	0.409	0.124	0.447	0.021
Arizona	0.712	0.127	0.001	0.160
Arkansas	0.015	0.950	0.010	0.025
California	0.690	0.258	0.039	0.013

Pour les qualités de représentation (cos2), les 3 fonctions apportent le même résultat.

Table 87: Extrait des Contributions avec PCA

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.783	2.596	1.107	0.282
Alaska	3.067	2.327	23.343	2.218
Arizona	2.507	1.124	0.017	8.034
Arkansas	0.016	2.534	0.074	0.385
California	5.137	4.811	2.010	1.349

Table 88: Extrait des Contributions avec prcomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.768	2.544	1.085	0.276
Alaska	3.005	2.281	22.876	2.174
Arizona	2.457	1.102	0.016	7.873
Arkansas	0.016	2.483	0.072	0.378
California	5.034	4.714	1.969	1.322

Table 89: Extrait des Contributions avec princomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.783	2.596	1.107	0.282
Alaska	3.067	2.327	23.343	2.218
Arizona	2.507	1.124	0.017	8.034
Arkansas	0.016	2.534	0.074	0.385
California	5.137	4.811	2.010	1.349

Et pour finir avec les contributions, les fonctions PCA et princomp ont la même sortie. tandis que prcomp propose des contributions différentes.

Exercice 32

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

Le but sera de représenter les éventuels liens entre la CSP et le type d'hébergement choisit HEB.

Table 90: Tableau de contingence de CSP et HEB

	CAMP	HOTEL	LOCA	RESI	TOTAL
AGRI	239	155	129	0	523
CADR	1003	1556	1821	1521	5901
INAC	682	1944	967	1333	4926
OUVR	2594	1124	2176	1038	6932
TOTAL	4518	4779	5093	3892	18282

Voici le tableau de contingence que nous utiliserons pour notre analyse.

```
##
##  Pearson's Chi-squared test
##
## data:  data[1:4, 1:4]
## X-squared = 2067.9, df = 9, p-value < 2.2e-16
```

On commence par réaliser un test du χ^2. On trouve statistique du χ^2 de 2067.9, et une pvaleur associée très proche de 0. Donc on rejette l'hypothèse d'indépendance, il y a un lien à étudier entre les CSP et HEB.

Table 91: Tableau des distributions conditionnelles des HEB sachant la CSP (%)

	CAMP	HOTEL	LOCA	RESI	TOTAL
AGRI	45.70	29.64	24.67	0.00	100
CADR	17.00	26.37	30.86	25.78	100
INAC	13.84	39.46	19.63	27.06	100
OUVR	37.42	16.21	31.39	14.97	100
TOTAL	24.71	26.14	27.86	21.29	100

On commence par les profils lignes. On apprend dans le tableau 91 que 37.42% des ouvriers choisissent le camping comme hébergement de vacance.

Table 92: Tableau des distributions conditionnelles des CSP sachant HEB (%)

	CAMP	HOTEL	LOCA	RESI	TOTAL
AGRI	5.29	3.24	2.53	0.00	2.86
CADR	22.20	32.56	35.75	39.08	32.28
INAC	15.10	40.68	18.99	34.25	26.94
OUVR	57.41	23.52	42.73	26.67	37.92
TOTAL	100.00	100.00	100.00	100.00	100.00

Avec les profils colonnes, on voit que parmi ceux qui choisissent une résidence secondaire comme logement de vacance, 39.08% sont des cadres.

On réalise ensuite l'AFC de nos données. On commence par étudiant l'inertie de la variance.

Table 93: Valeur propre

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.098	86.855	86.855
Dim.2	0.014	12.256	99.111
Dim.3	0.001	0.889	100.000

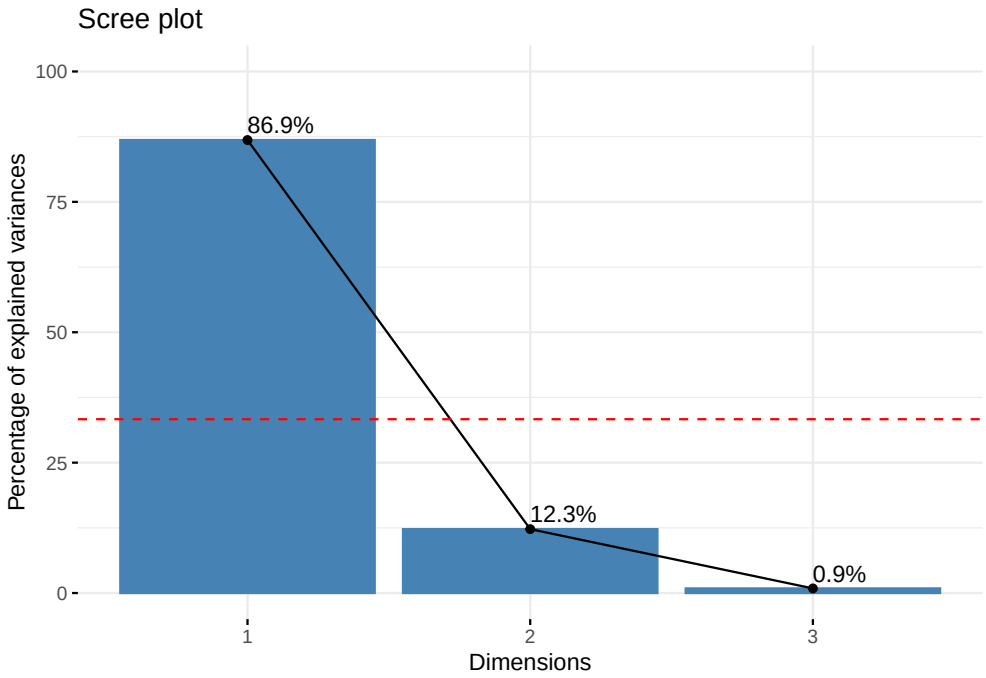

Figure 25: Visualisation des valeurs propres

Voici le tableau et le graphique de nos valeurs propres. Quand on fait la somme des valeurs propres multipliées par n, on retrouve bien la statistique du χ^2.

les deux premiers axes expliquent 99.9% de la variance totale. C'est quasiment la totalité. Les dimensions 1 et 2 expliquent environ 86,5% et 12.256% de l'inertie totale, respectivement. On conserve ces 2 dimensions.

On trouve ensuite les différents indicateurs et commence par les modalités de CSP.

Table 94: Coordonnées des modalités de CSP

	Dim 1	Dim 2	Dim 3
AGRI	-0.441	-0.431	-0.137
CADR	0.140	0.129	-0.027
INAC	0.379	-0.109	0.020
OUVR	-0.355	0.001	0.019

Le tableau 94 nous montre les coordonnées que prendront les modalités de CSP sur les graphiques pour chaque dimension.

Table 95: Cos2 des modalités de CSP

	Dim 1	Dim 2	Dim 3
AGRI	0.488	0.465	0.047
CADR	0.532	0.449	0.019
INAC	0.921	0.077	0.003
OUVR	0.997	0.000	0.003

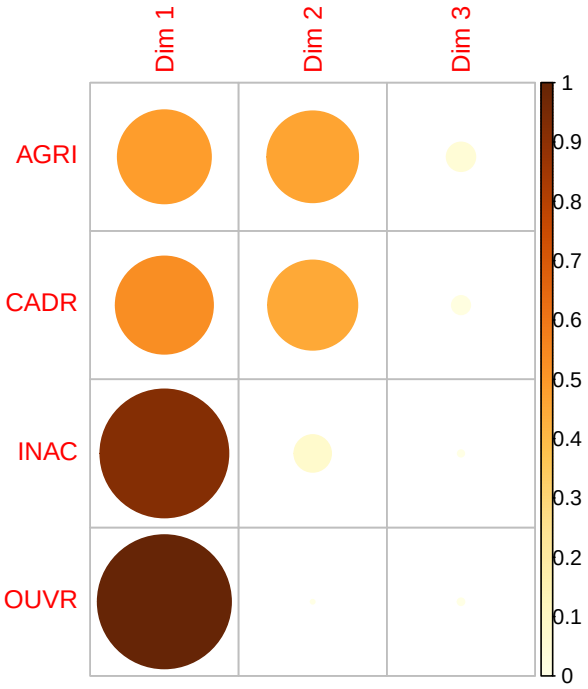

Figure 26: Visualisation des cos2 des modalités de CSP

On s'intéresse ensuite aux qualités de représentation. On remarque dans le tableau et sur le graphique, que les inactifs et les ouvriers auront une bonne représentation sur l'axe 1.

Table 96: Contributions des modalités de CSP

	Dim 1	Dim 2	Dim 3
AGRI	5.676	38.347	53.116
CADR	6.430	38.451	22.841
INAC	39.307	23.200	10.548
OUVR	48.586	0.002	13.495

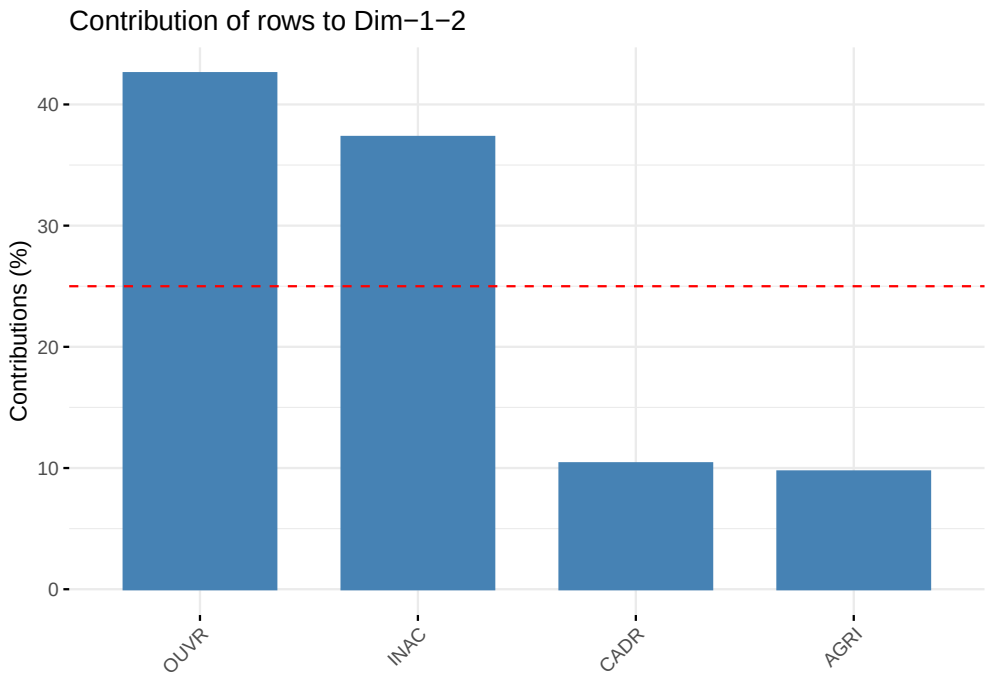

Figure 27: Visualisation des contributions des modalités de CSP

Regardons maintenant les contributions. Le tableau 96 nous montre le pourcentage de la contribution pour chaque axe. Avec le graphique, on peut voir que les modalités OUVR et INAC sont les plus importantes dans la définition de le premier plan. La ligne pointillée rouge correspond à la contribution moyenne.

Maintenant passons aux modalités de HEB.

Table 97: Coordonnées des modalités de HEB

	Dim 1	Dim 2	Dim 3
CAMP	-0.443	-0.088	0.022
HOTEL	0.325	-0.139	-0.019
LOCA	-0.130	0.124	-0.036
RESI	0.286	0.110	0.045

Le tableau 97 nous montre les coordonnées que prendront les modalités de HEB sur les graphiques pour chaque dimension.

Table 98: Cos2 des modalités de HEB

	Dim 1	Dim 2	Dim 3
CAMP	0.960	0.038	0.002
HOTEL	0.842	0.155	0.003
LOCA	0.504	0.457	0.039
RESI	0.852	0.127	0.021

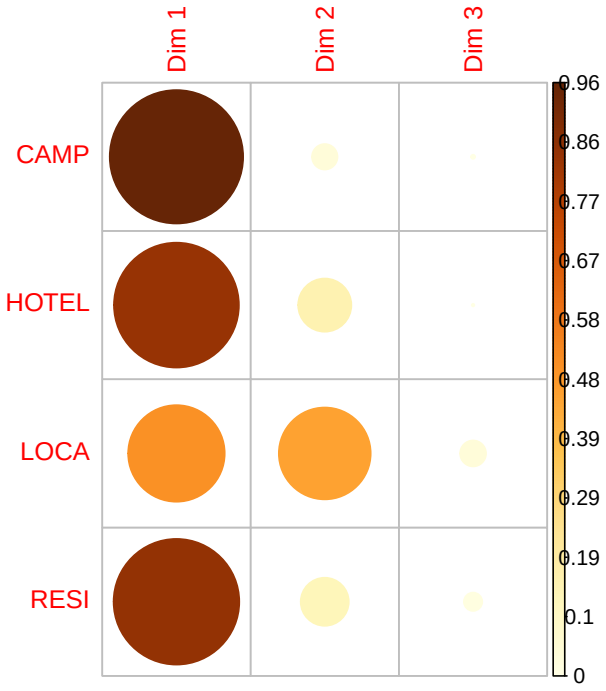

Figure 28: Visualisation des cos2

On s'interesse ensuite aux qualités de représentation. On remarque dans le tableau et sur le graphique, que les campings et les hôtels auront une bonne représentation sur la dimension 1.

Table 99: Contributions des modalités de HEB

	Dim 1	Dim 2	Dim 3
CAMP	49.372	13.714	12.201
HOTEL	28.056	36.594	9.210
LOCA	4.822	30.953	36.367
RESI	17.750	18.739	42.222

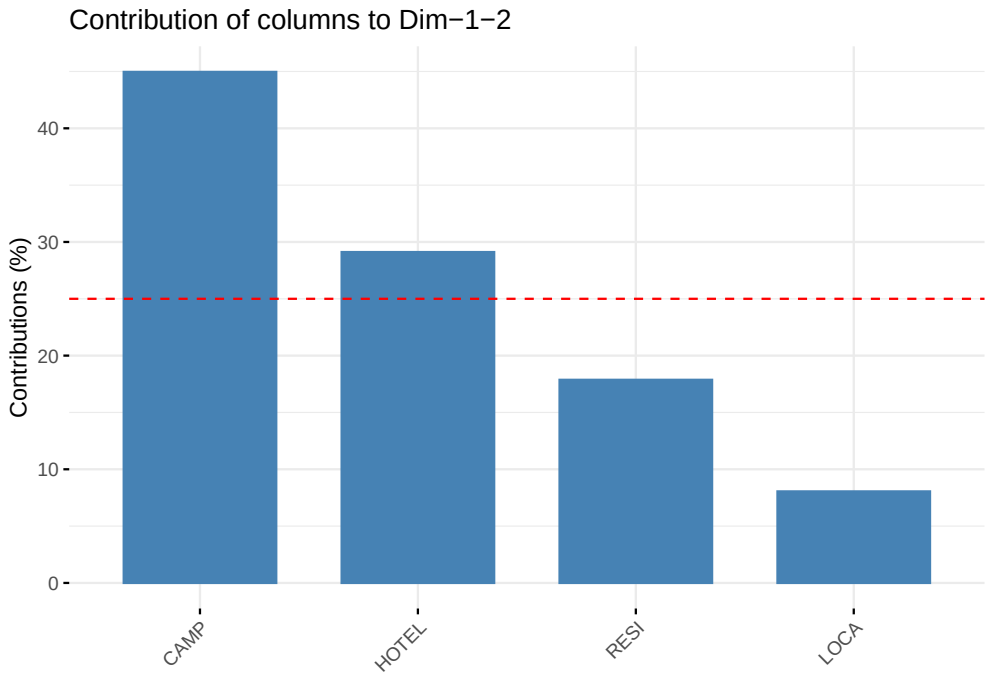

Figure 29: Visualisation des contributions des modalités de HEB

Regardons maintenant les contributions. Le tableau nous montre le pourcentage de la contribution pour chaque axe. Avec le graphique, on peut voir que les modalités CAMP et HOTEL sont les plus importantes dans la définition du premier plan. La ligne pointillée rouge correspond à la contribution moyenne.

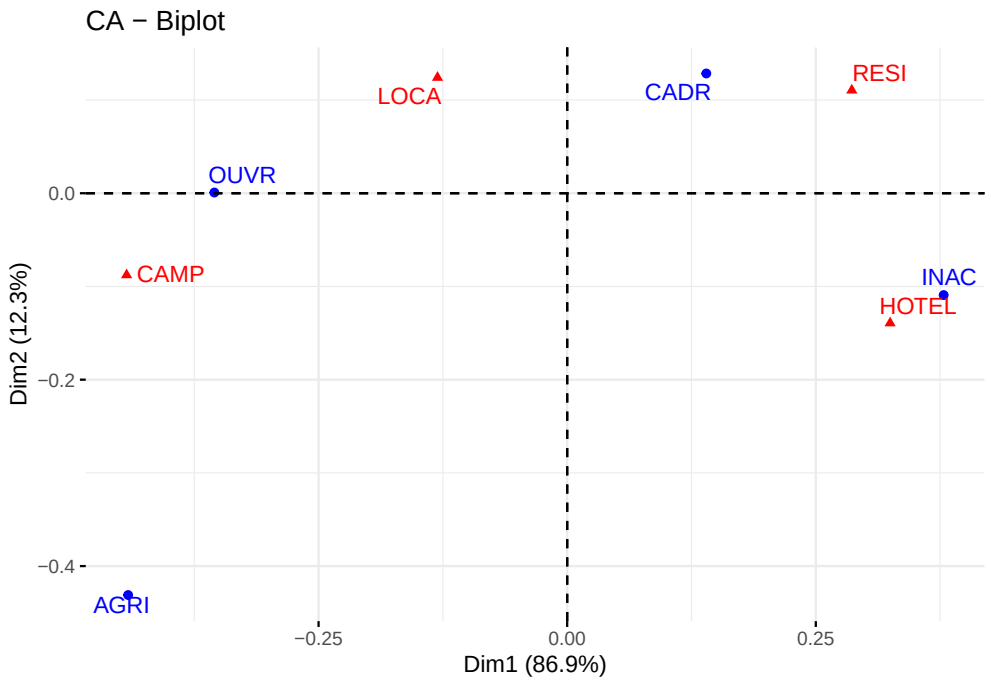

Figure 30: Bitplot

On peut ensuite tracer le biplot entre nos deux variables sur le premier plan. Les modalités de CSP sont représentées par des points bleus et les modalités de HEB par des triangles rouges.

Quand on s'intéresse uniquement aux modalités de CSP, on remarque les cadres et agriculteurs sont opposés, ce qui indique que leurs profils s'oppose également.

Pour les modalités de HEB, on trouve ce phénomène entre les résidences secondaires et les campings.

La forme générale est un arc de cercle il y a donc un effet de Guttman. Il y a donc un ordre de nos modalités. On retrouve des groupes, on voit que les inactifs sont liés aux hôtels, les cadres sont plus proches des résidences secondaires, et les ouvriers opteront plus pour des campings. On peut imaginer qu'il y a un lien avec le coût de ces types d'Herbergement.

Exercice 33

Dans cette partie, nous analyserons un tableau de contingence donnant les fréquences de 4 catégories de fumeur (en colonne) pour 5 catégories de salariés (en ligne) dans une entreprise fictive. Les catégories en ligne sont :

— SM=Senior Managers,

— JM=Junior Managers,

— SE=Senior Employees,

— JE=Junior Employees,

— SC=Secretaries.

Table 100: Tableau de contingence de nos données

	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

Voici les données smokes que nous utiliserons.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 101: Tableau de contingence avec marge

	none	light	medium	heavy	sum
SM	4	2	3	2	11
JM	4	3	7	4	18
SE	25	10	12	4	51
JE	18	24	33	13	88
SC	10	6	7	2	25
sum	61	45	62	25	193

On peut ajouter les distributions marginales.

Table 102: Tableau de contingence en fréquence (%)

	none	light	medium	heavy	sum
SM	2.073	1.036	1.554	1.036	5.699
JM	2.073	1.554	3.627	2.073	9.326
SE	12.953	5.181	6.218	2.073	26.425
JE	9.326	12.435	17.098	6.736	45.596
SC	5.181	3.109	3.627	1.036	12.953
sum	31.606	23.316	32.124	12.953	100.000

On peut mettre le tableau de nos données en pourcentage.

Table 103: Tableau des effectifs théoriques

	none	light	medium	heavy
SM	3.476684	2.564767	3.533679	1.424870
JM	5.689119	4.196891	5.782383	2.331606
SE	16.119171	11.891192	16.383420	6.606218
JE	27.813471	20.518135	28.269430	11.398964
SC	7.901554	5.829015	8.031088	3.238342

Et aussi calculer le tableau des effectifs théoriques, utiles pour le test du $\chi 2$.

Table 104: Profils lignes

	none	light	medium	heavy	sum
SM	36.364	18.182	27.273	18.182	100
JM	22.222	16.667	38.889	22.222	100
SE	49.020	19.608	23.529	7.843	100
JE	20.455	27.273	37.500	14.773	100
SC	40.000	24.000	28.000	8.000	100
sum	31.606	23.316	32.124	12.953	100

Table 105: Profils colonnes

	none	light	medium	heavy	sum
SM	6.557	4.444	4.839	8	5.699
JM	6.557	6.667	11.290	16	9.326
SE	40.984	22.222	19.355	16	26.425
JE	29.508	53.333	53.226	52	45.596
SC	16.393	13.333	11.290	8	12.953
sum	100.000	100.000	100.000	100	100.000

Voici les tableaux des profils lignes et colonnes. Avec les profils lignes on voit que 28% des secretaries fument “moyennement”. Avec les profils colonnes, on remarque que parmi ceux qui ne fument pas, 40.98% sont Senior Employees.

On réalise ensuite l'AFC sur nos données.

Table 106: Valeurs propres

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.075	87.756	87.756
Dim.2	0.010	11.759	99.515
Dim.3	0.000	0.485	100.000

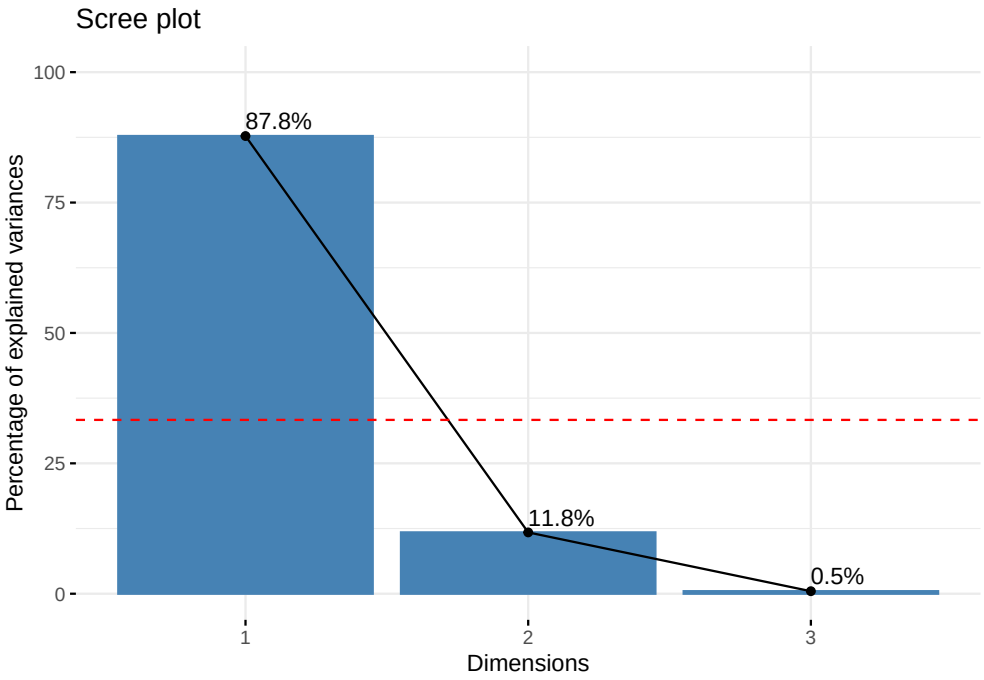

Figure 31: Visualisation des valeurs propres

On commence par les valeurs propres. On voit dans le tableau 106 et la figure 31 que le premier plan explique 99.5% de la variance totale. Les dimensions 1 et 2 expliquent environ 87.7% et 11.7% de l'inertie totale, respectivement. On conserve ces 2 dimensions.

Table 107: Coordonnées pour les catégories de salarié

	Dim 1	Dim 2	Dim 3
SM	-0.066	0.194	0.071
JM	0.259	0.243	-0.034
SE	-0.381	0.011	-0.005
JE	0.233	-0.058	0.003
SC	-0.201	-0.079	-0.008

Table 108: Coordonnées pour les catégories de fumeur

	Dim 1	Dim 2	Dim 3
none	-0.393	0.030	-0.001
light	0.099	-0.141	0.022
medium	0.196	-0.007	-0.026
heavy	0.294	0.198	0.026

On récupère ensuite les coordonnées pour tracer le bitplot. D'abord pour les catégories de salariés, puis pour les catégories de fumeur.

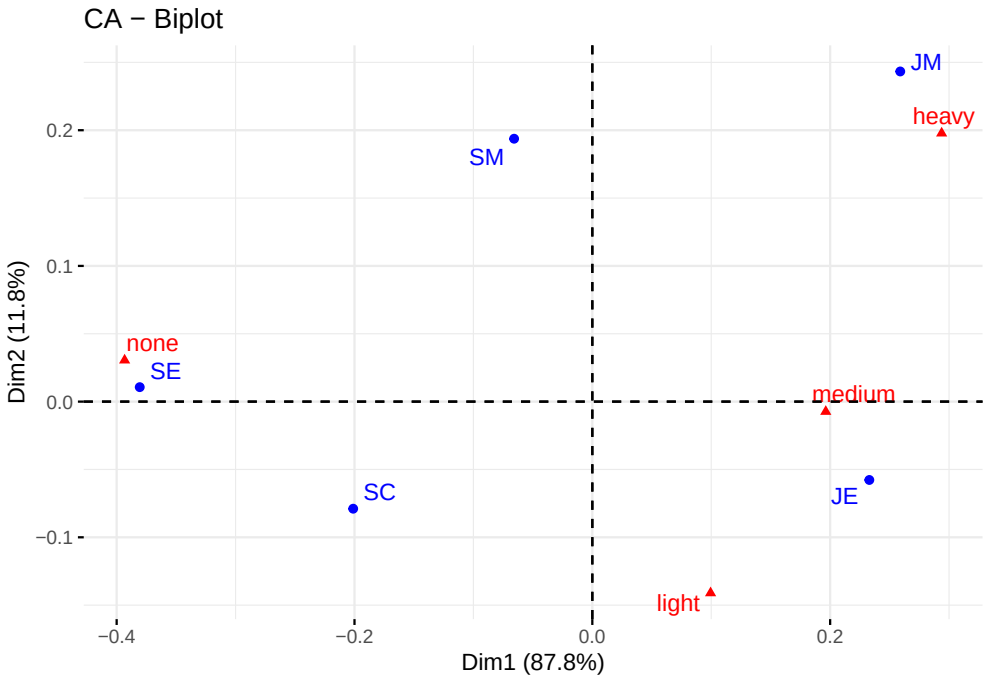

Figure 32: bitplot

Et on trace ensuite le bitplot. Les catégories de salariés sont représentées par des points bleus et les catégories de fumeur par des triangles rouges.

On remarque qu'il n'y a pas de groupe qui se forme entre catégories d'une même variable, par contre pour les catégories de salariés on voit qu'il y a une opposition entre les secretaries et les Juniors managers, ce qui signifait que leurs profils s'opposent également.

Quand on regarde les deux variables ensembles on voit des regroupements. Par exemple on se rend compte du lien qu'il y a entre les Senior Employees et les non-fumeurs, aussi entre ce qui fume "moyennement" et les Juniors employees, et entre les gros fumeurs et les juniors managers. Nous avons réussi à bien cibler les liens qui existent entre la catégorie de salariés et les catégories de fumeur.

Exercice 34

Il s'agit ici de proposer une méthodologie d'analyse textuelle pour identifier les auteurs de deux fragments de texte anonymes. On connaît pour chacun de ces fragments de texte la fréquence d'apparition de certaines lettres. On suppose également 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 15 échantillons est donnée dans un tableau de contingence.

Nous réaliserons l'AFC, puis nous recommencerons avec deux textes supplémentaires ou l'auteur n'est pas spécifié.

```
##
##  Pearson's Chi-squared test
##
## data:  ecrivain[1:15]
## X-squared = 533.46, df = 224, p-value < 2.2e-16
```

On commence par voir s'il y a indépendance des données. On remarque que la pvaleur est proche de zéro, donc on rejette l'hypothèse d'indépendance. L'AFC est légitime.

Table 109: Tableau des valeurs propres

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.018	36.037	36.037
Dim.2	0.010	18.967	55.004
Dim.3	0.008	14.996	70.000
Dim.4	0.005	10.603	80.603
Dim.5	0.004	7.072	87.675
Dim.6	0.002	4.173	91.848

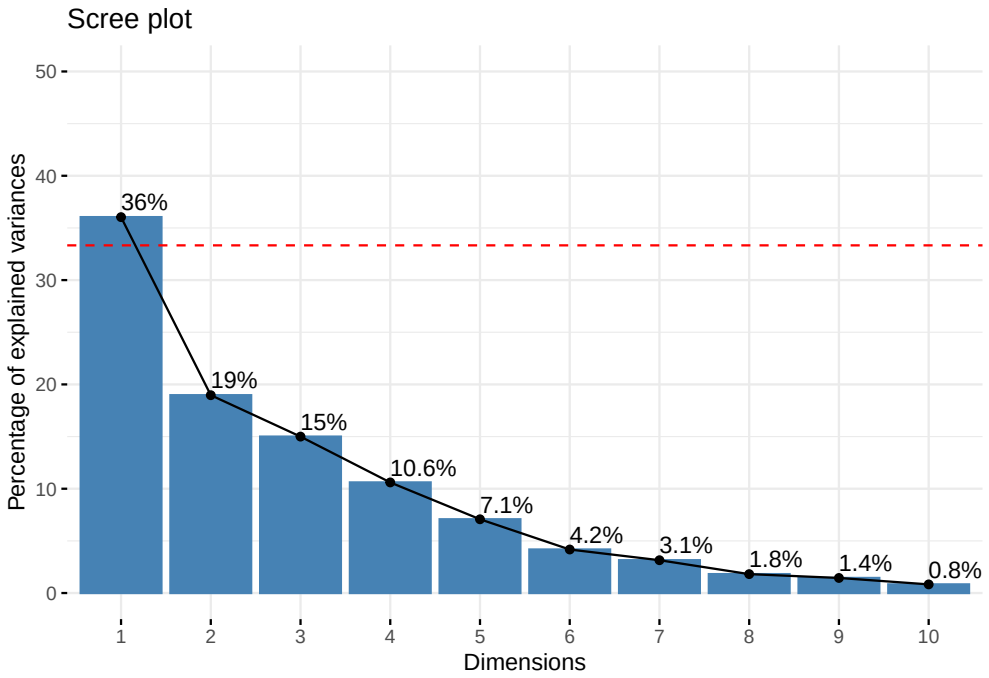

Figure 33: bitplot

On commence par déterminer le nombre d'axes. Avec le graphique et le tableau on remarque que les quatre premiers axes expliquent 80.6% de la variance totale. C'est un pourcentage acceptable. On conservera donc 4 axes dans notre analyse. Les résultats sont assez similaires quand on ajoute les individus supplémentaires, on conservera 4 axes également.

Table 110: Extrait des cos2 des auteurs

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
CD1	0.092	0.593	0.101	0.138	0.008
CD2	0.127	0.310	0.078	0.127	0.033
CD3	0.279	0.211	0.396	0.040	0.053
RD1	0.028	0.058	0.631	0.192	0.000
RD2	0.164	0.104	0.173	0.002	0.344
RD3	0.404	0.086	0.239	0.029	0.043

Table 111: Extrait des cos2 des lettres

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
B	0.028	0.269	0.618	0.023	0.021
C	0.394	0.464	0.003	0.008	0.031
D	0.667	0.053	0.011	0.007	0.152
F	0.216	0.012	0.255	0.218	0.171
G	0.205	0.136	0.445	0.067	0.000
H	0.000	0.453	0.115	0.229	0.100

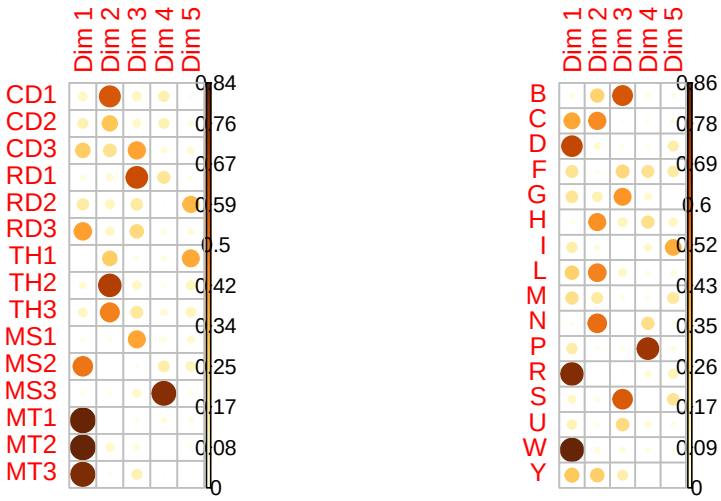

Figure 34: Visualisation des cos2 des auteurs et des lettres

On s'intéresse maintenant à la qualité de représentation (cos2). On remarque avec le graphique et le tableau que les textes de Mark Twain ont la meilleure représentation sur l'axe 1. Pour les lettres, ce sont le R et W qui sont bien représenté sur l'axe 1, on peut aussi voir les contributions sur les autres axes.

Table 112: Extrait des contributions des auteurs

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
CD1	2.022	24.868	5.332	10.376	0.925
CD2	1.562	7.238	2.304	5.324	2.049
CD3	8.458	12.130	28.787	4.137	8.108
RD1	0.296	1.152	15.966	6.880	0.007
RD2	1.910	2.305	4.836	0.096	20.400
RD3	5.178	2.085	7.367	1.252	2.811

Table 113: Extrait des contributions des lettres

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
B	0.706	12.760	37.104	1.995	2.666
C	8.945	20.030	0.180	0.638	3.642
D	15.370	2.329	0.611	0.540	17.829
F	3.414	0.363	9.698	11.716	13.842
G	2.636	3.317	13.708	2.925	0.026
H	0.001	12.952	4.159	11.714	7.697

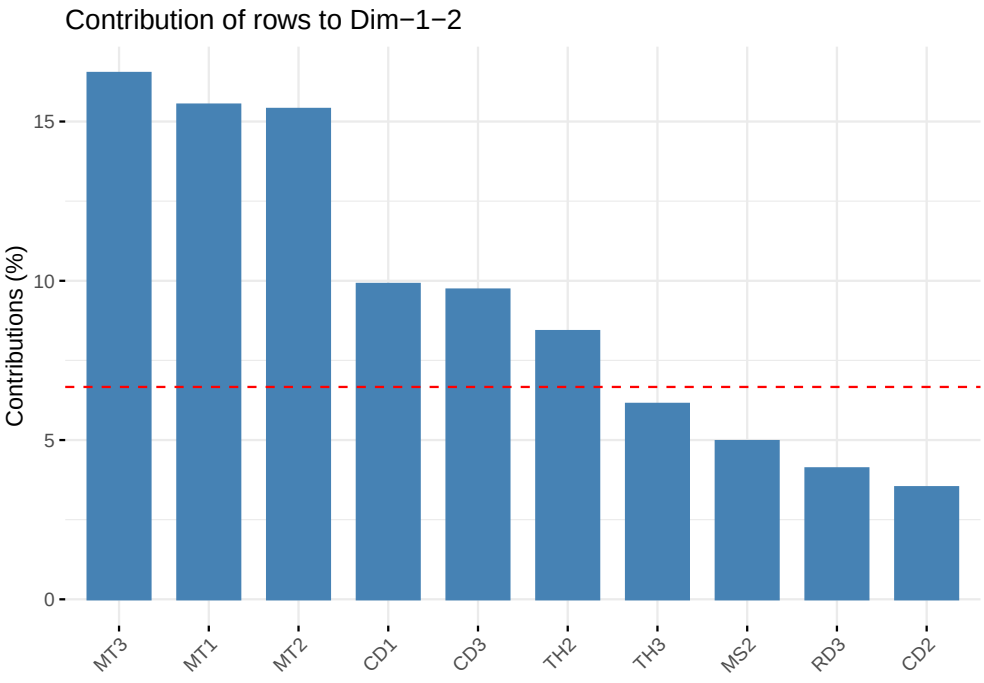

Figure 35: Visualisation des contributions des auteurs

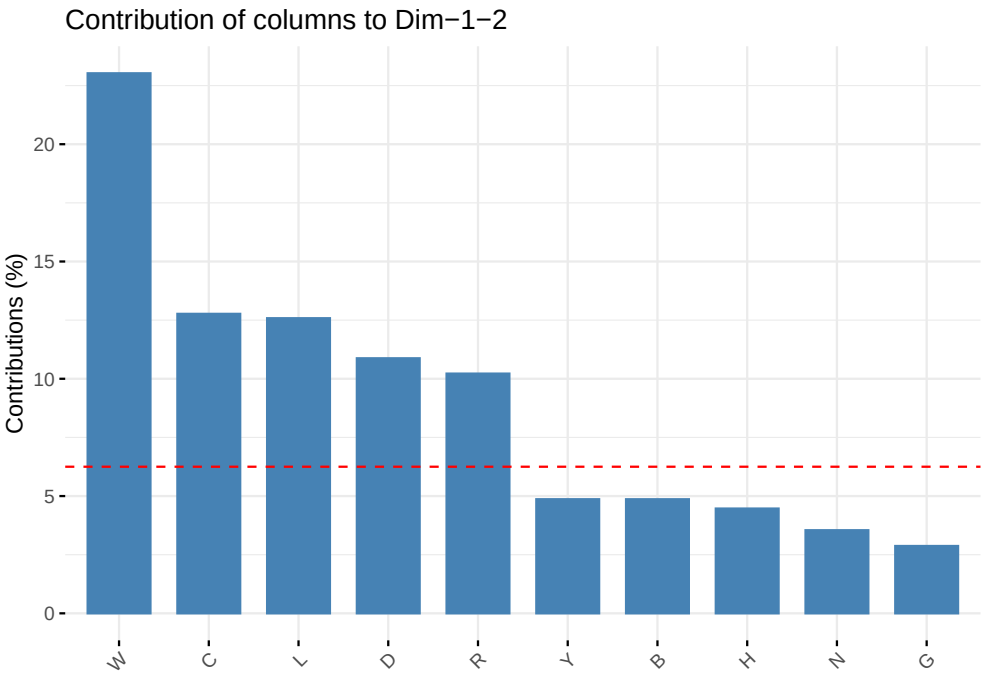

Figure 36: Visualisation des contributions des lettres

On regarde maintenant les contributions sur le premier plan, avec les graphiques et les tableaux ci-dessus. On voit que les auteurs qui contribuent le plus au premier plan sont Mark Twain et Charles Darwin. Pour les lettres on voit que ce sont les lettres W,C,L,D,R qui contribuent le plus au premier plan.

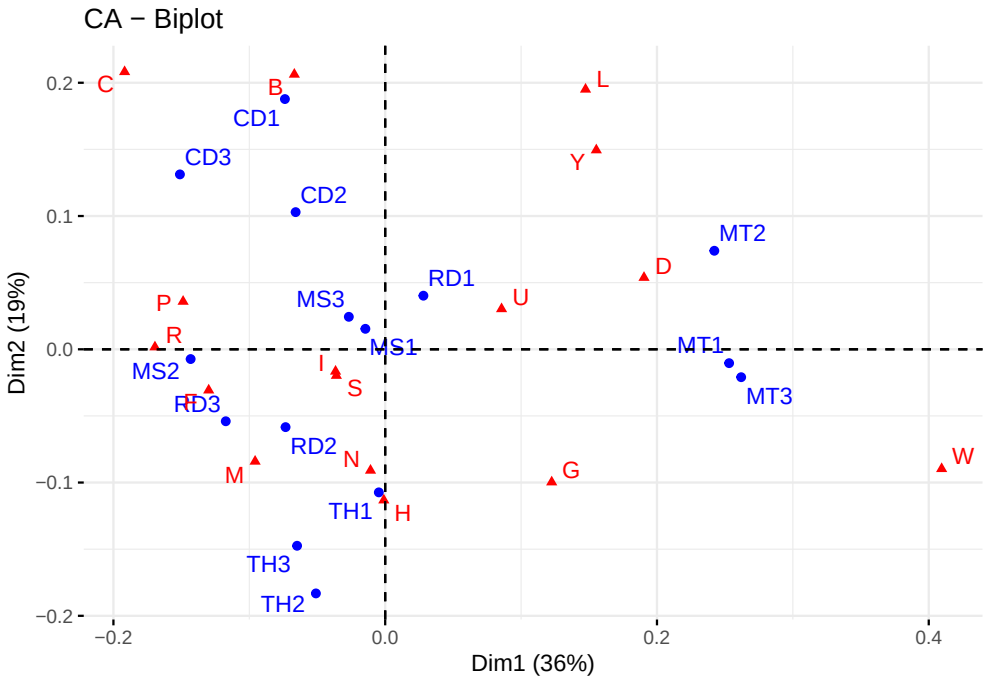

Figure 37: Bitplot

Quand on trace le bit plot, on remarque qu'il y a des groupes d'auteurs qui se forment, associer à certaines lettres. Par exemple pour Charles Darwin, on voit que les lettres B et C sont celles où il y a le plus de lien. Les auteurs René Descartes et Mary Shelley sont très liés, il semble avoir la même utilisation de lettre. Mark Twain se distingue des autres, il est lié à la lettre D et aussi le plus proche du W. Pour finir Thomas Hobbes se distingue aussi mais de façon moins prononcée, il se confond presque avec René Descartes et Mary Shelley.

On réalise ensuite l'AFC avec les textes supplémentaires. On trouve des contributions et des qualités de représentation très similaires aux résultats précédents.

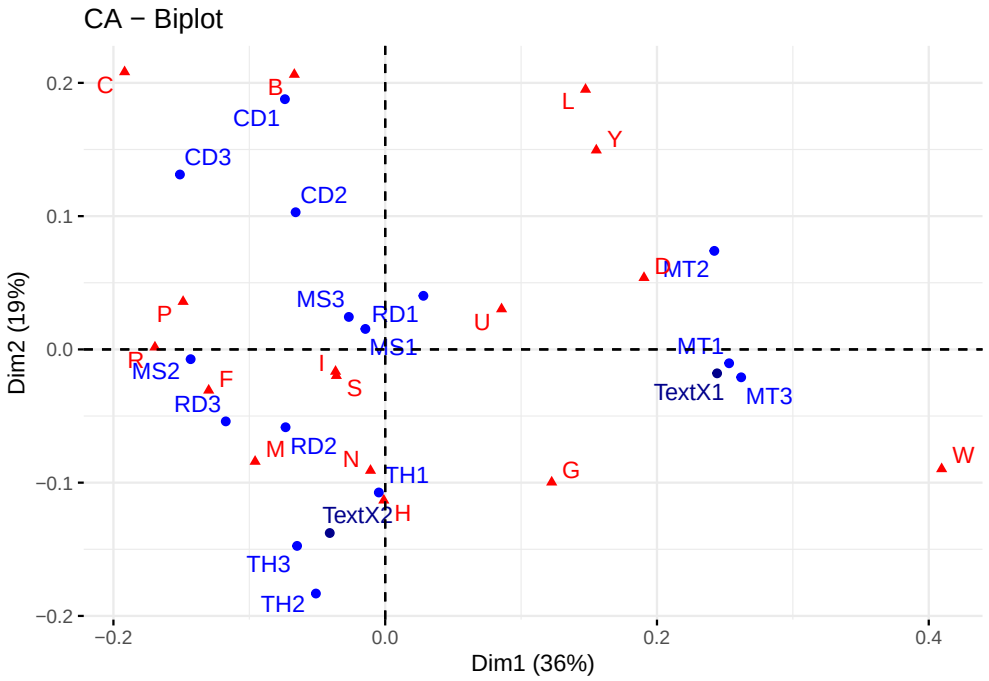

Figure 38: Bitplot avec les textes supplémentaire

On peut refaire un bitplot en incluant c'est deux textes. On voit que le texte un est fortement similaire à un texte de Mark Twain, alors que le texte 2 semble plus lié à un texte de Thomas Hobbes. Nous allons classifier nos textes pour voir si cela se confirme.

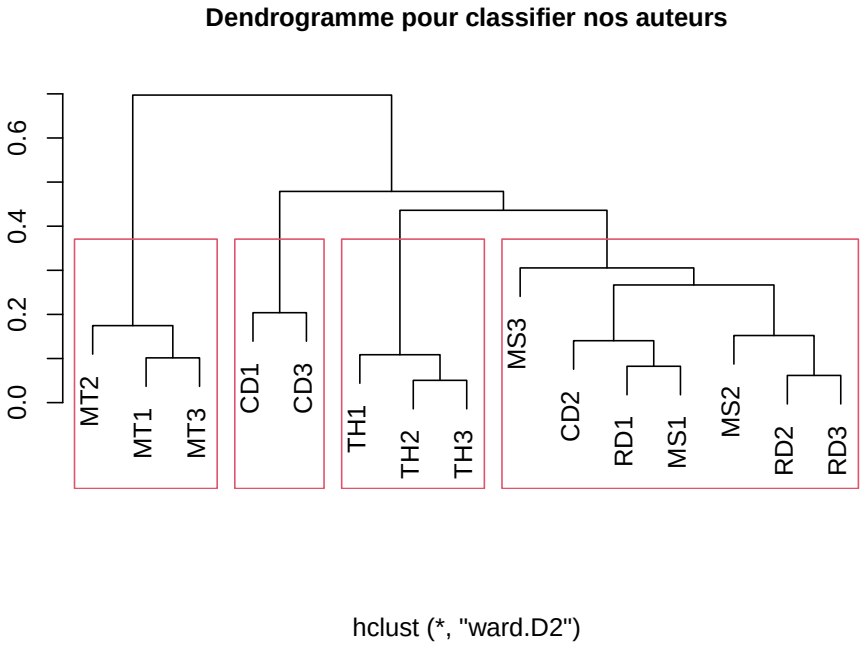

Figure 39: Dendrogramme pour classifier nos auteur

Le dendrogramme suivant nous montre la partition en 4 classes que nous offrent nos données. Cela nous permet de confirmer les conclusions que nous avons tirée avec les bitplots. On voit qu'une classe est composé des textes de t Mark Twain avec le texte 1 qui doit aussi être un texte de cet auteur.

La deuxième classe comporte les textes de Thomas Hobbes avec les textes 2 qui doit être issu de cet auteur. La troisième classe est composée de deux textes de Charles Darwin, notre classification n'a pas considéré le troisième texte de Charles Darwin dans cette classe. En effet ce texte se trouve dans la quatrième classe, avec les textes de Mary Shelley et René Descartes, qui comme nous l'avons dit ont tendance à utiliser les mêmes lettres.

Chapitre 5 : Analyse Factorielle des Correspondances Multiples (AFM)

Exercice 27

Nous traiterons des données fictives ou 27 races de chiens sont décrites avec 7 variables qualitatives.

Table 114: Extrait des données chiens

	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

Voici un extrait des données, nous avons 6 variables ordinales, la taille, le poids , la vélocité, l'intelligence, l'affectation et l'agressivité, et une variable fonction qui détermine l'utilité des chiens, qui peut être utilité, chasse ou compagnie. Pour notre analyse nous ne conserverons que les variables ordinales.

Nous allons réaliser un AFM, avec fonction comme variable supplémentaire.

Table 115: Valeurs propres

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.482	28.896	28.896
Dim.2	0.385	23.084	51.981
Dim.3	0.211	12.657	64.638
Dim.4	0.158	9.453	74.091
Dim.5	0.150	9.008	83.099
Dim.6	0.123	7.398	90.497
Dim.7	0.081	4.888	95.385
Dim.8	0.046	2.740	98.125
Dim.9	0.024	1.413	99.537
Dim.10	0.008	0.463	100.000

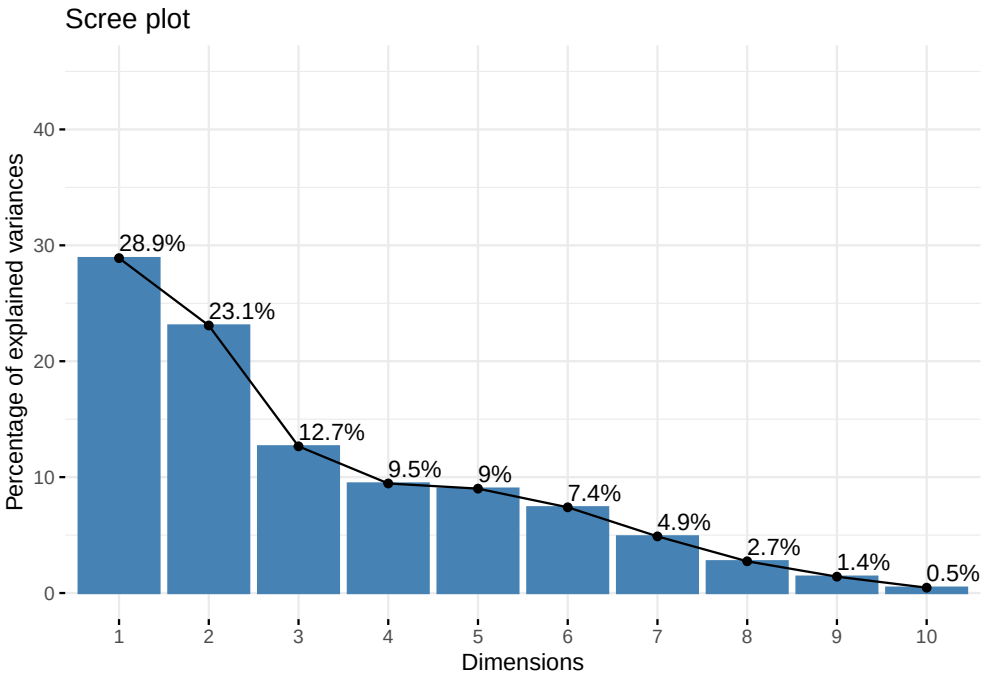

Figure 40: Visualisation des valeurs propres

On commence par les valeurs propres, on voit avec le tableau qu'à partir de 4 dimensions, plus de 70% de l'inertie totale, on conserve donc les 4 premiers axes. Avec le graphique ont voit que l'axe 1 explique 28.9%, l'axe 2 23.1%, l'axe 3 12.7% et l'axe 4 9.5%.

Table 116: Extrait des coordonnée des variables

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
T-	1.185	0.924	-0.616	0.120	-0.020
T+	0.851	-1.232	1.016	0.342	-0.310
T++	-0.837	-0.021	-0.051	-0.170	0.113
P-	1.169	0.824	-0.359	0.165	-0.051
P+	-0.305	-0.819	-0.231	-0.118	-0.190
P++	-1.015	0.974	1.222	0.068	0.615

On se focalise d'abord sur les variables. On obtient dans le tableau ci-dessus les coordonnées afin de tracer le grapahique des variables.

Table 117: Extrait des Cos2 des variables

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
T-	0.491	0.299	0.133	0.005	0.000
T+	0.165	0.345	0.235	0.027	0.022
T++	0.875	0.001	0.003	0.036	0.016
P-	0.575	0.286	0.054	0.011	0.001
P+	0.100	0.722	0.058	0.015	0.039
P++	0.234	0.216	0.339	0.001	0.086

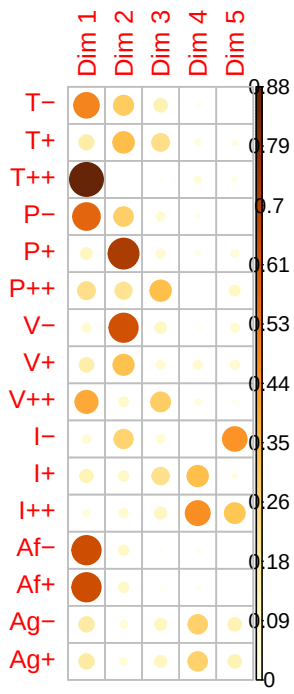

Figure 41: Visualisation des cos2 des variables

On se penche ensuite sur les qualités de représentations (cos2) des variables. On voit à l'aide du graphique et du tableau qu'une grande taille aura une bonne qualité de représentation sur l'axe 1.

Table 118: Extrait des contributions des variables

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
T-	12.598	9.587	7.772	0.396	0.011
T+	4.642	12.171	15.104	2.297	1.976
T++	13.459	0.010	0.115	1.703	0.783
P-	14.010	8.722	3.013	0.852	0.086
P+	1.674	15.062	2.191	0.768	2.082
P++	6.604	7.609	21.833	0.090	7.763

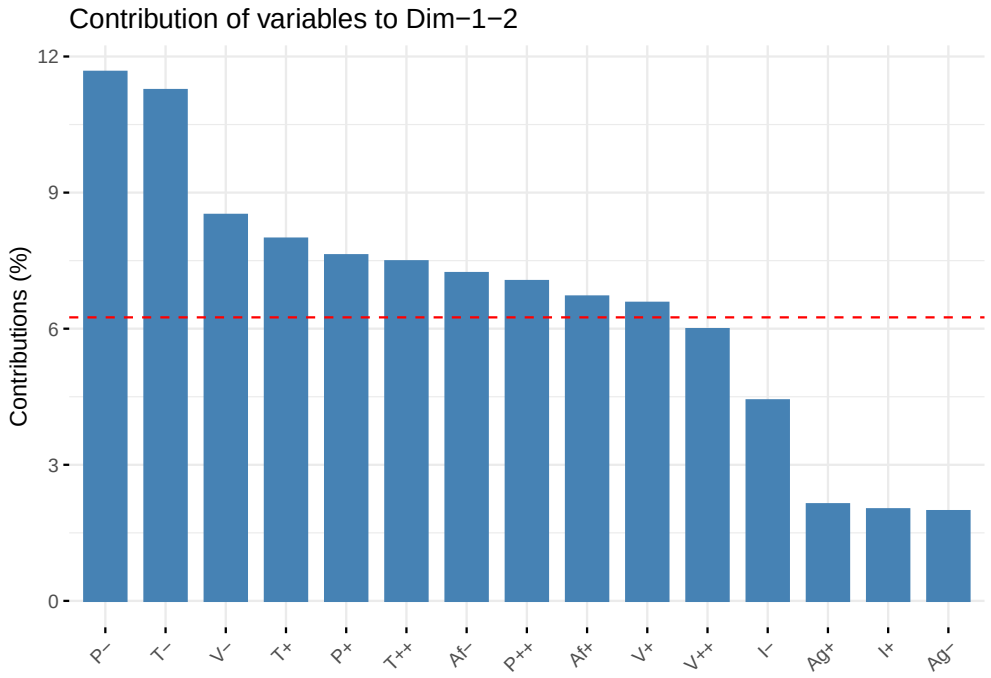

Figure 42: Visualisation des contributions pour les variables

Pour les contributions des variables sur le premier plan. On remarque avec le graphique et le tableau qu'un faible poids à la meilleure contribution au premier plan. Sur le graphique toutes les variables au-dessus de la ligne pointillée rouge peuvent être considéré comme suffisamment contribuant au premier plan.

On passe maintenant aux individus.

Table 119: Extrait des coordonnées des individus

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
beauceron	-0.317	-0.418	-0.101	-0.211	-0.119
basset	0.254	1.101	-0.191	0.293	-0.524
ber_allem	-0.486	-0.464	-0.498	0.577	0.276
boxer	0.447	-0.882	0.692	0.260	-0.456
bull-dog	1.013	0.550	-0.163	-0.350	0.331
bull-mass	-0.753	0.547	0.498	0.655	0.722

D'abord avec le tableau des coordonnées.

Table 120: Extrait des Cos2 des individus

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
beauceron	0.089	0.154	0.009	0.039	0.012
basset	0.034	0.635	0.019	0.045	0.144
ber_allem	0.154	0.140	0.161	0.217	0.049
boxer	0.111	0.433	0.266	0.038	0.115
bull-dog	0.624	0.184	0.016	0.074	0.067
bull-mass	0.271	0.143	0.118	0.205	0.249

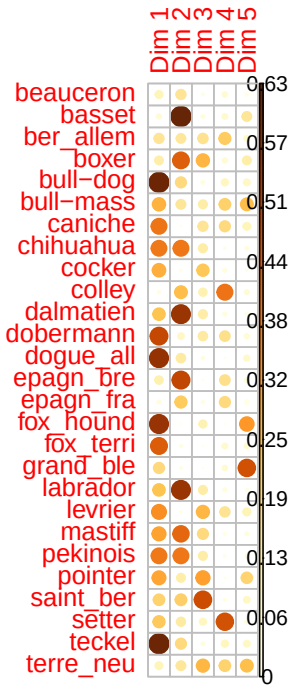

Figure 43: Visualisation des cos2 des individus

Ensuite avec les qualités de représentations des individus. On ne voit par exemple avec le graphique que les teckels et les bull-dogs ont la meilleure qualité de représentation sur la dimension 1.

Table 121: Extrait des contributions des individus

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
beauceron	0.774	1.680	0.181	1.051	0.346
basset	0.497	11.674	0.638	2.013	6.774
ber_allem	1.819	2.077	4.357	7.838	1.878
boxer	1.539	7.485	8.408	1.589	5.120
bull-dog	7.897	2.911	0.469	2.878	2.699
bull-mass	4.356	2.879	4.347	10.090	12.858

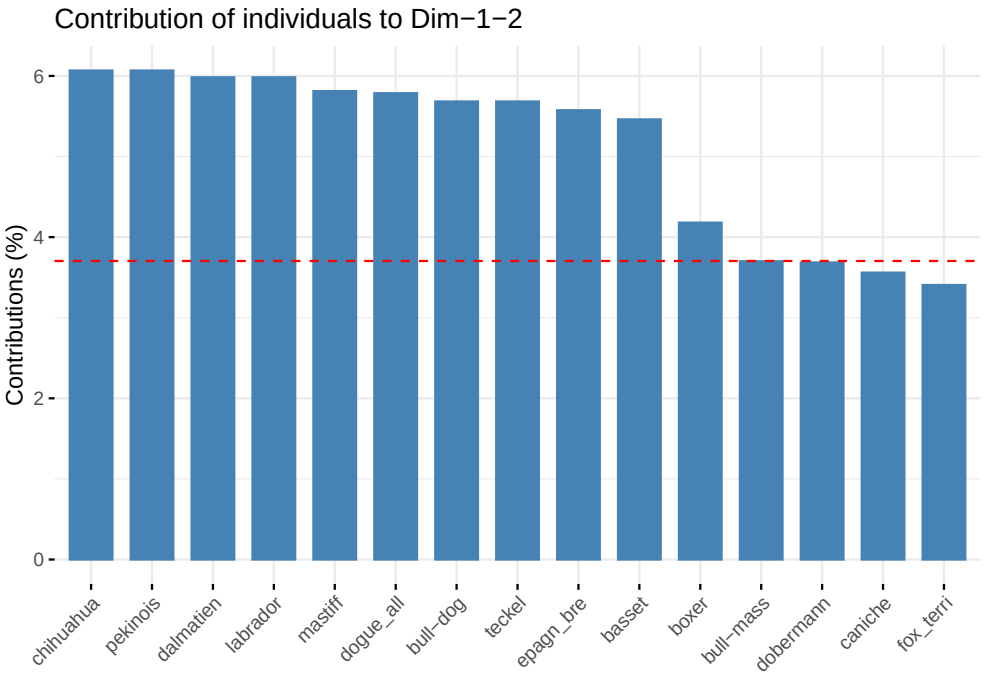

Figure 44: Visualisation des contributions des individus

Pour les contributions des individus sur le premier plan. On remarque avec le graphique les chihuahuas ont la meilleure contribution au premier plan. Tous les individus au-dessus de la ligne pointillée rouge peuvent être considéré comme suffisamment contribuant au premier plan.

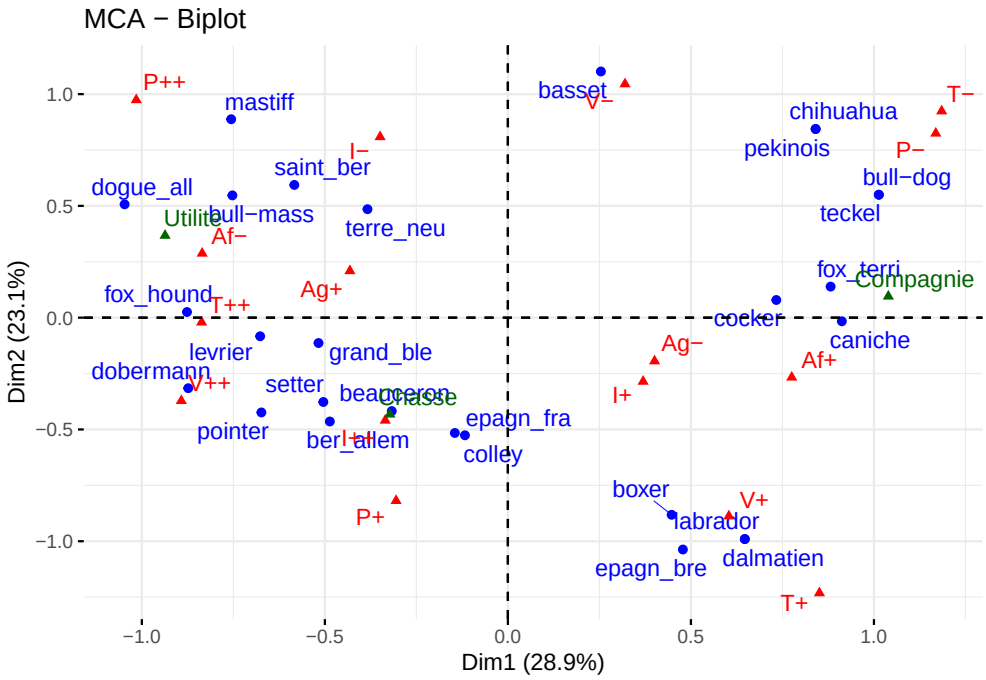

Figure 45: Bitplot

On peut enfin tracer le bitplot. Les individus sont en bleu, les variables sont en rouge et les variables supplémentaires sont en vert foncé.

On peut faire des liens entre les individus et les variables, tous les individus proches les uns des autres peuvent être considérés comme des profils similaires. par exemple on voit que les boxers, les labradors, les dalmatiens et les espagn_bre sont similaires avec une grande taille et une vitesse élevée.

Quand on regarde les variables supplémentaires, on voit qu'elles sont éloigner les unes des autres surtout pour les chiens de compagnie qu'on arrive bien à distinguer des deux autres.

On va s'intéresser aux rapports de corrélations entre les variables qualitatives et les deux premières composantes principales.

Table 122: Rapports de corrélations entre les variables qualitatives et les deux premières composantes principales

	Dim 1	Dim 2
taille	0.887	0.502
poids	0.644	0.725
velocite	0.411	0.684
intellig	0.127	0.280
affect	0.648	0.077
agress	0.173	0.041

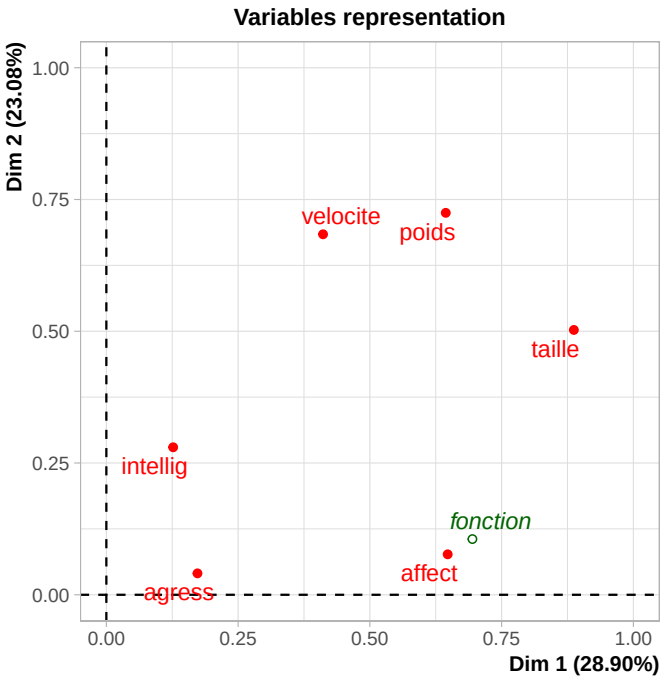

Figure 46: Visualisation des rapports de corrélation

On voit avec le tableau et le graphique que le poids est la variable la plus corrélée à l'axe 1 tandis que le poids et la plus corrélée à l'axe 2.

On décide ensuite de rajouter des données manquantes à nos données, et nous refaisons une AFM, pour voir si elles sont prises en compte.

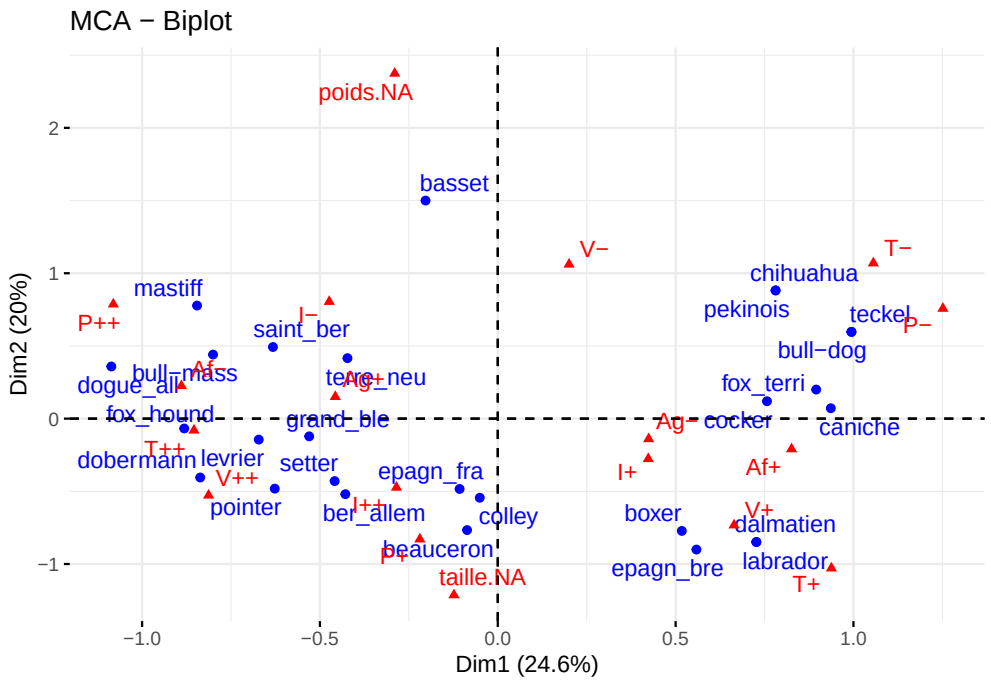

Figure 47: Bitplot avec les données manquantes

Quand on refait le bitplot on voit bien des points supplémentaires avec -NA en suffixe, donc les données manquantes sont prises en compte par la fonction MCA comme des individus classiques, ce qui n'est pas correct.

On veut maintenant comparer l'ACM et l'AFC dans le cas particulier de deux variables qualitatives. Nous allons réaliser l'AFC du tableau de contingence croisant les variables taille et poids, et comparer les valeurs propres.

Table 123: Valeurs propre de l'AFC

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.861	91.743	91.743
Dim.2	0.077	8.257	100.000

Table 124: Valeurs propre de l'AFM

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.964	48.193	48.193
Dim.2	0.639	31.958	80.151
Dim.3	0.361	18.042	98.193
Dim.4	0.036	1.807	100.000

Table 125: Valeurs propres AFM avec l'AFC

	Valeurs propres
dim 1	0.964
dim 2	0.639
dim 4	0.036
dim 3	0.361

On retrouve un lien entre les valeurs propres de l'AFC et l'ACM. Quand on joue avec la racine carrée des valeurs propres de l'AFC, on arrive à retouver les valeurs propres de l'ACM.

$\frac{1+\sqrt{vpDim1AFC}}{2} = vp\ dim1\ ACM$

$\frac{1+\sqrt{vpDim2AFC}}{2} = vp\ dim2\ ACM$

$\frac{1-\sqrt{vpDim1AFC}}{2} = vp\ dim3\ ACM$

$\frac{1-\sqrt{vpDim2AFC}}{2} = vp\ dim4\ ACM$

Ou vp sont les valeurs propres selon la méthode et la dimension.

Exercice 28

Dans cette partie, nous allons présenter le package R missMDA. Il gère les données manquantes en ACP et en ACM, et de choisir le nombre de composantes par validation croisée. Nous décrirons les principales fonctionnalités de ce package, avec à chaque fois une explication de la méthode.

Overimpute : Évaluez l'ajustement de la distribution prédictive après avoir effectué une imputation multiple

estim_ncpPCA : Estime le nombre de dimensions pour l'Analyse en Composantes Principales par validation croisée

MIFAMD : Effectue des imputations multiples pour des données mixtes (continues et catégorielles) en utilisant l'analyse factorielle de données mixtes.

estim_ncpMultilevel : Estimez le nombre de dimensions pour la composante principale multiniveau (ACP multiniveau, AMC multiniveau ou analyse factorielle multiniveau de données mixtes) par validation croisée.

estim_ncpMCA : Estimer le nombre de dimensions pour l'Analyse des Correspondances Multiples par validation croisée

MIPCA : Réalise une imputation multiple avec un modèle ACP. Peut être utilisé comme étape préliminaire pour effectuer une imputation multiple dans l'ACP.

MIMCA : Effectue des imputations multiples pour des données catégorielles en utilisant l'analyse des correspondances multiples.

estim_ncpFAMD : Estime le nombre de dimensions pour l'Analyse Factorielle de Données Mixtes par validation croisée

prelim : Cette fonction effectue des opérations de regroupement et de tri sur un ensemble de données imputées à plusieurs reprises. Elle crée un objet mids qui est nécessaire à l'entrée de with.mids, qui permet d'analyser l'ensemble de données imputées à plusieurs reprises. L'ensemble de données incomplètes d'origine doit être disponible pour que nous sachions où se trouvent les données manquantes.

imputeFAMD : Imputez les valeurs manquantes d'un ensemble de données mixtes (avec des variables continues et catégorielles) en utilisant la méthode des composantes principales "analyse factorielle pour données mixtes" (FAMD). Peut être utilisé comme une étape préliminaire avant d'exécuter FAMD sur un ensemble de données incomplet.

imputeMFA : Impute un jeu de données avec des variables structurées en groupes de variables (groupes de variables continues ou catégorielles).

imputeMCA : Imputez les valeurs manquantes d'un ensemble de données catégoriques en utilisant l'analyse des correspondances multiples (ACM). Peut être utilisé comme une étape préliminaire avant d'effectuer l'ACM sur un ensemble de données incomplet.

imputeCA : Imputez les entrées manquantes d'un tableau de contingence en utilisant l'analyse des correspondances (AC). Peut être utilisé comme une étape préliminaire avant d'effectuer l'AC sur un ensemble de données incomplet.

imputePCA : Impute les valeurs manquantes d'un jeu de données avec le modèle d'analyse en composantes principales. Peut être utilisé comme une étape préliminaire avant d'effectuer une ACP sur un jeu de données complet.

imputeMultilevel : Imputez les valeurs manquantes d'un ensemble de données mixtes multi-niveaux (avec une variable qui regroupe les individus, et avec des variables continues et catégorielles) en utilisant la méthode des composantes principales "analyse factorielle multi-niveaux pour données mixtes".

plot.MIMCA : À partir des ensembles de données imputées multiples, la fonction trace des graphiques pour les individus, les catégories et les dimensions pour l'analyse des correspondances multiples (ACM).

plot.MIPCA : À partir des ensembles de données imputées multiples, la fonction trace des graphiques pour les individus, les variables et les dimensions pour l'analyse en composantes principales (ACP).

Projet personnel : CSP et la principale source d'information

Nous étudions la relation entre la catégorie socio-professionnelle (CSP) et la principale source d'information sur les problèmes d'environnement. Sept CSP sont étudiées :

agriculteur (AGRI),

cadre supérieur (CSUP),

cadre moyen (CMOY),

employé (EMPL),

ouvrier (OUVR),

retraité (RETR),

chômeur (CHOM).

Les 1283 personnes interrogées devaient indiquer leur principale source d'information sur les problèmes d'environnement, parmi les six sources suivantes :

télévision (TEL),

journaux (JOU),

radio (RAD),

livres (LIV),

associations (ASS) ,

mairie (MAI).

Table 126: Tableau de contingence

	TEL	JOU	RAD	LIV	ASS	MAI	Total
AGRI	26	18	9	5	4	6	68
CSUP	19	49	4	16	5	3	96
CMOY	44	87	4	39	14	3	191
EMPL	83	87	13	24	5	1	213
OUVR	181	107	16	31	7	7	349
RETR	167	95	29	15	7	7	320
CHOM	27	9	4	2	2	2	46
Total	547	452	79	132	44	29	1283

Voici le jeu de données à notre disposition pour menez notre analyse. Il s'agit d'un tableau des effectifs croisés entre les deux variables que nous étudions. On apprend par exemple que 181 ouvriers ont comme principale source d'information sur les problèmes d'environnement la télévision.

Indépendance

Avant de commencer à analyser nos données on vérifie s'il y a un lien entre nos deux variables. Pour cela on réalise un test d'indépendance du χ^2.

```
##
##  Pearson's Chi-squared test
##
## data:  media_sm
## X-squared = 156.33, df = 30, p-value < 2.2e-16
```

On trouve une pvaleur très proche de 0, on rejette donc l'hypothèse d'indépendance, il y a un lien entre la catégorie socio-professionnelle et la principale source d'information sur les problèmes d'environnement. Une AFC est donc légitime.

Analyse Factorielle des Correspondances

Nous allons donc réaliser l'AFC sur nos données.

Valeurs propres

On cherche d'abord combien de dimension nous allons garder pour représenter au mieux nos données.

Table 127: Valeurs propres

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.092	75.099	75.099
Dim.2	0.022	18.001	93.100
Dim.3	0.006	5.269	98.369
Dim.4	0.002	1.437	99.806
Dim.5	0.000	0.194	100.000

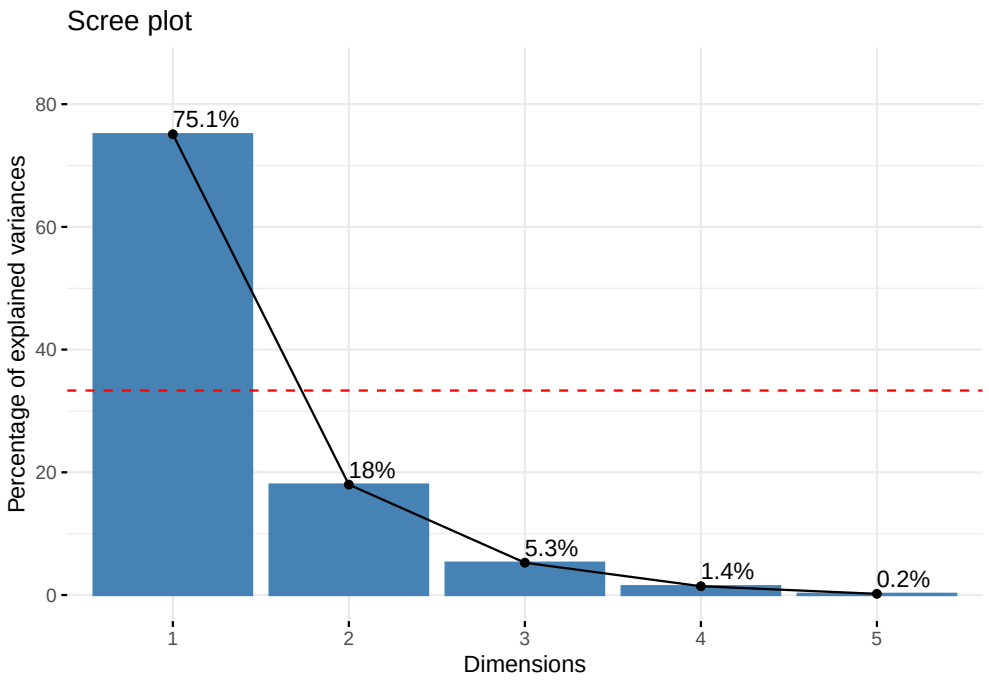

Figure 48: Visulation des valeurs prorpes

Dans le tableau on voit que les deux premiers axes expliquent 93.1% de la variance totale. C'est un pourcentage très acceptable. Avec le graphique on voit que les dimensions 1 et 2 expliquent environ 75.% et 18% de l'inertie totale, respectivement. On décide donc de retenir ces deux axes.

Catégories socio-professionnelle

On commence par étudier la variable CSP.

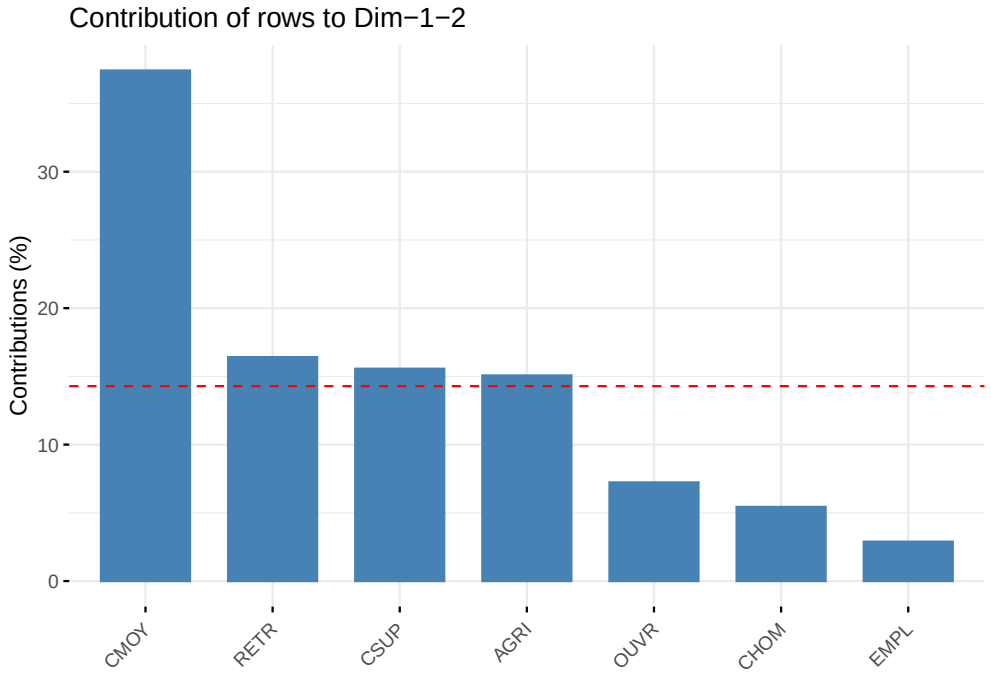

Figure 49: Visulation des contrivutions des CSP sur le premier plan

On cherche à savoir quelle catégorie socio-professionnelle contribue le plus au premier plan. Avec le diagramme en bar ci-dessus, on voit les différentes contributions pour chaque catégorie. La ligne pointillée rouge correspond à la contribution moyenne, toutes les variables au-dessus de cette ligne peuvent être considérées comme fortement contribuent au premier plan. On retient donc les agriculteurs, la classe supérieure, les retraités, et la classe moyenne qui a la meilleure contribution au premier plan.

Table 128: Coordonnées des CSP sur le premier plan

	Dim 1	Dim 2
AGRI	-0.150	0.548
CSUP	0.477	0.093
CMOY	0.534	0.015
EMPL	0.085	-0.112
OUVR	-0.145	-0.096
RETR	-0.273	0.005
CHOM	-0.388	0.147

Avant de représenter le nuage des CSP, on regarde les coordonnées de chaque catégorie sur le premier plan. Tous les individus avec une coordonnée négative sur la dimension 1, seront dans la partie gauche du graphique, et inversement pour les coordonnées positive. Et tous les individus avec une coordonnée négative sur la dimension 2, seront dans la partie basse du graphique, et inversement pour les coordonnées positive. Par exemple les agriculteurs seront dans la partie haute gauche du graphique.

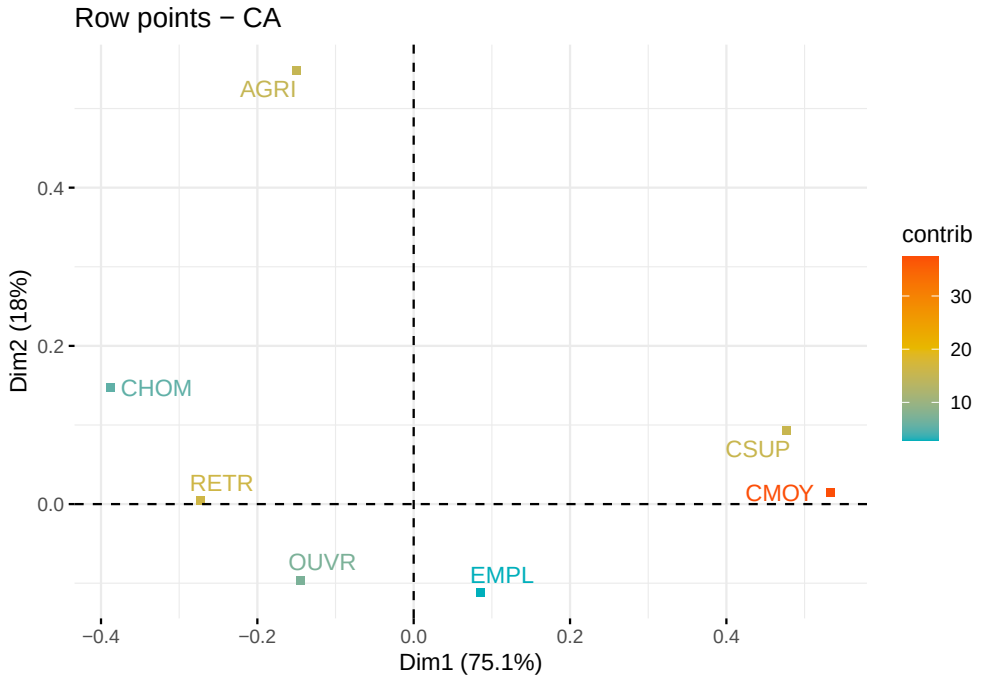

Figure 50: Nuage des CSP sur le premier plan

On peut ensuite tracer le nuage des catégories socio-professionnelles. Ici il y a une couleur selon la contribution, plus la contribution au premier plan et grande, plus la couleur sera chaude. On retrouve donc la classe moyenne avec la couleur la plus chaude car comme nous venons de le voir c'est la catégorie qui contribue le plus à ce plan.

Ici on remarque que la classe moyenne et la classe supérieure sont regroupées, cela signifie qu'ils ont un profil similaire. À l'inverse les catégories qui s'opposent vont être corrélées négativement. Par exemple les employés et les agriculteurs ont ici des profils qui s'opposent mais cela reste léger.

Sources d'information

On s'intéresse maintenant à la deuxième variable de nos données.

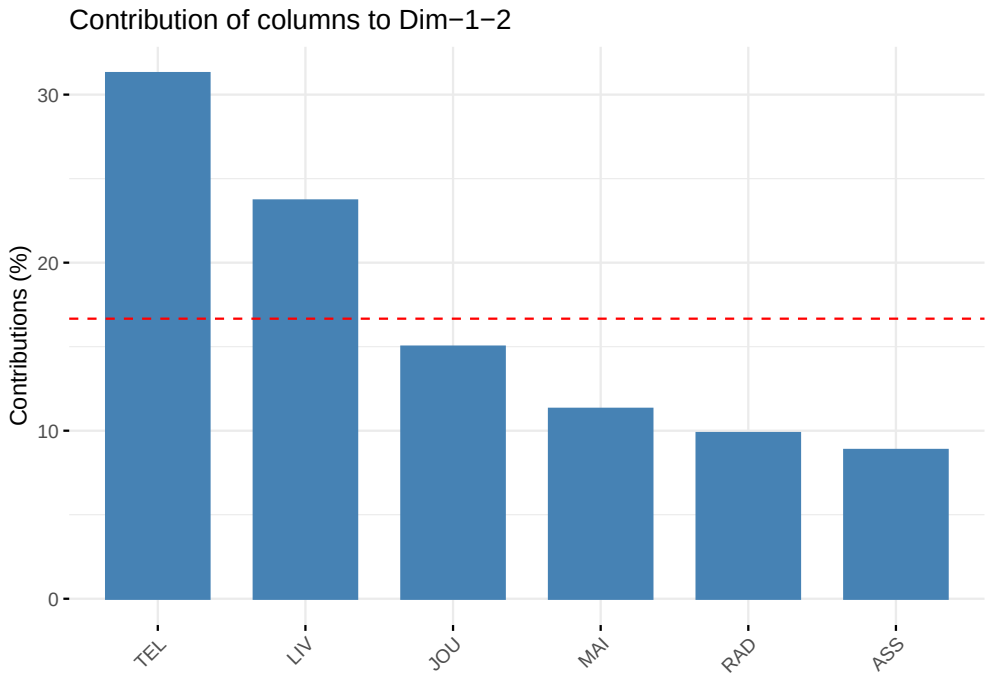

Figure 51: Visulation des contributions des sources d'information

On commence par les contributions sur le premier plan. On voit avec le graphique que deux sources d'information ont une contribution supérieure à la contribution moyenne. Ce sont la télévision et les livres, ce seront donc ces deux modalités qui contribueront le plus au premier plan.

Table 129: Coordonnées des sources d'information sur le premier plan

	Dim 1	Dim 2
TEL	-0.281	-0.065
JOU	0.217	-0.037
RAD	-0.334	0.265
LIV	0.511	-0.025
ASS	0.450	0.301
MAI	-0.169	0.734

Avant de représenter le nuage des sources d'information, on regarde les coordonnées de chaque source sur le premier plan. Tous les individus avec une coordonnée négative sur la dimension 1, seront dans la partie gauche du graphique, et inversement pour les coordonnées positive. Et tous les individus avec une coordonnée négative sur la dimension 2, seront dans la partie basse du graphique, et inversement pour les coordonnées positive. Par exemple la télévision sera dans la partie basse gauche du nuage.

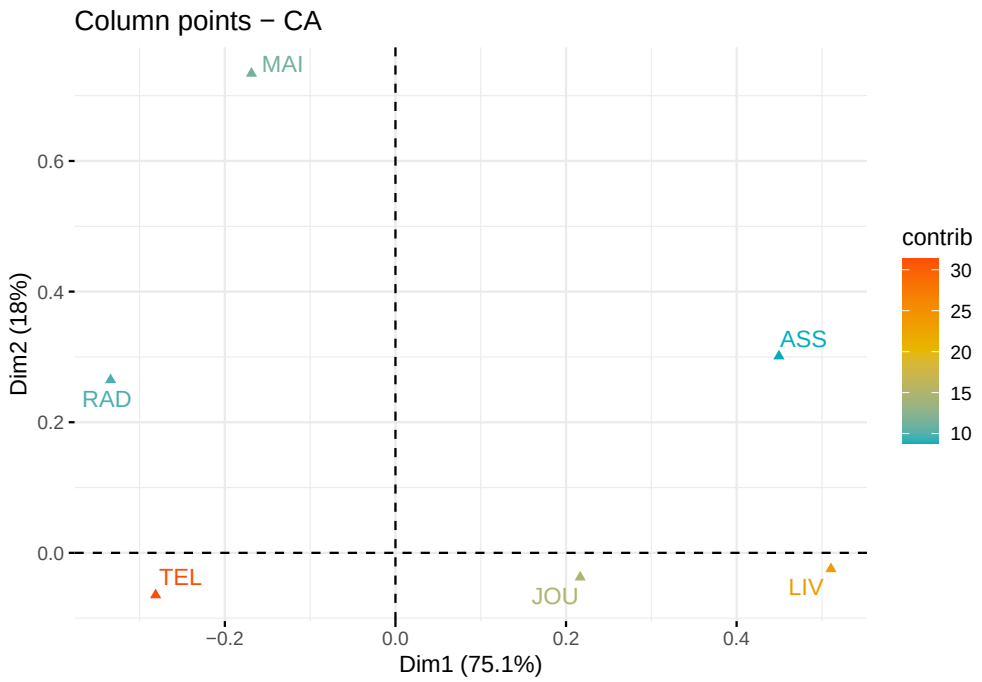

Figure 52: Nuage des sources d'informations

On peut ensuite tracer le nuage des sources d'information. Comme pour le nuage précédent, il y a une couleur selon la contribution, plus la contribution au premier plan est grande, plus la couleur sera chaude. On retrouve donc la télévision et les livres avec les contributions les plus élevés.

Ici il n'y a pas de groupe qui se forme, donc aucune source d'information n'a un profil similaire. Il y a quelque opposition, comme la télévision et les associations, donc les profils de ses deux sources d'information s'opposent, mais cela n'est pas net.

CSP et sources d'information

On peut maintenant croiser les deux variables pour voir les liens et les dissimilarités.

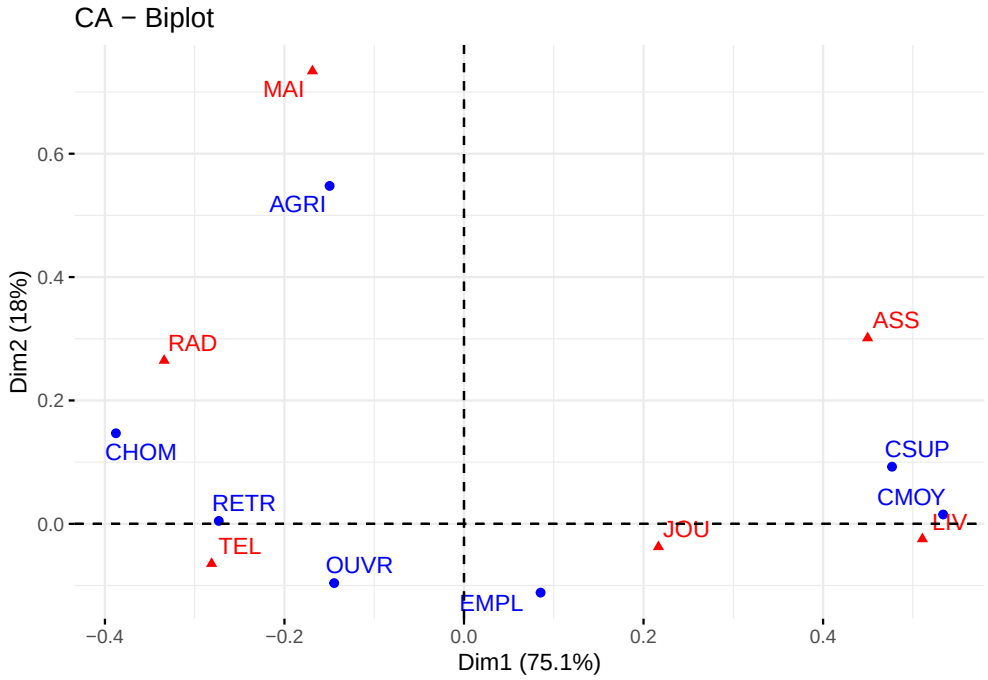

Figure 53: Graphe superposé

Pour cela on superpose les deux nuages. Les CSP sont représentées par des points bleus et des sources d'information par des triangles rouges.

On remarque que les modalités sont disposées en arc de cercle. Ce phénomène est connu sous le nom d'effet Guttman. Il y a donc un ordre sous-tend les modalités. On voit que la classe moyenne et la classe supérieure sont fortement liées aux livres, les employés aux journaux, les ouvriers et retraités à la télévision, les chômeurs aux radios, et les agriculteurs aux mairies.

Pour aller plus loin, on imagine avec ces résultats que le coût d'une source d'information à un impact. En effet on remarque que les sources d'information gratuite comme la télévision, la radio et la mairie, vont être associées aux catégories sociales les plus modestes. Tandis que les livres et les journaux, des sources d'information payante vont plutôt être liées aux catégories les plus aisés.