Chapitre 4 : Analyse Factorielle des Correspondances (AFC)

Exercice 31

Nous travaillons avec le jeu de données USArrests disponible dans R. Ces données contient 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 vas être de comparer les sortie de l'ACP sur ces données avec 3 fonctions différentes, PCA,prcomp, et princomp.

Table 1: Extrait des Coordonée avec PCA

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

Table 2: Extrait des Coordonée avec prcomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	-0.9756604	1.1220012	-0.4398037	0.1546966
Alaska	-1.9305379	1.0624269	2.0195003	-0.4341755
Arizona	-1.7454429	-0.7384595	0.0542302	-0.8262642
Arkansas	0.1399989	1.1085423	0.1134222	-0.1809736
California	-2.4986128	-1.5274267	0.5925410	-0.3385592

Table 3: Extrait des Coordonée avec princomp

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

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

Table 4: Extrait des Cos2 avec PCA

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.3920310	0.5184533	0.0796601	0.0098556
Alaska	0.4085425	0.1237310	0.4470626	0.0206638
Arizona	0.7122238	0.1274849	0.0006875	0.1596038
Arkansas	0.0151456	0.9496046	0.0099411	0.0253086
California	0.6904652	0.2580267	0.0388312	0.0126769

Table 5: Extrait des Cos2 avec prcomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.3920310	0.5184533	0.0796601	0.0098556
Alaska	0.4085425	0.1237310	0.4470626	0.0206638
Arizona	0.7122238	0.1274849	0.0006875	0.1596038
Arkansas	0.0151456	0.9496046	0.0099411	0.0253086
California	0.6904652	0.2580267	0.0388312	0.0126769

Table 6: Extrait des Cos2 avec princomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.3920310	0.5184533	0.0796601	0.0098556
Alaska	0.4085425	0.1237310	0.4470626	0.0206638
Arizona	0.7122238	0.1274849	0.0006875	0.1596038
Arkansas	0.0151456	0.9496046	0.0099411	0.0253086
California	0.6904652	0.2580267	0.0388312	0.0126769

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

Table 7: Extrait des Contribution avec PCA

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.7832625	2.595723	1.1070956	0.2816054
Alaska	3.0666668	2.327394	23.3429239	2.2182476
Arizona	2.5068088	1.124411	0.0168326	8.0337329
Arkansas	0.0161272	2.533823	0.0736314	0.3853982
California	5.1369800	4.810526	2.0095751	1.3488039

Table 8: Extrait des Contribution avec prcomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.7675973	2.543809	1.0849536	0.2759732
Alaska	3.0053335	2.280846	22.8760654	2.1738826
Arizona	2.4566726	1.101923	0.0164959	7.8730583
Arkansas	0.0158047	2.483147	0.0721588	0.3776903
California	5.0342404	4.714315	1.9693836	1.3218278

Table 9: Extrait des Contribution avec princomp

	Dim.1	Dim.2	Dim.3	Dim.4
Alabama	0.7832625	2.595723	1.1070956	0.2816054
Alaska	3.0666668	2.327394	23.3429239	2.2182476
Arizona	2.5068088	1.124411	0.0168326	8.0337329
Arkansas	0.0161272	2.533823	0.0736314	0.3853982
California	5.1369800	4.810526	2.0095751	1.3488039

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-professionnel (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 choisi HEB.

Table 10: 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 khi2. On trouve statistique du khi-deux de 2067.9, et une pvaleur associés très proche de 0. Donc on rejette l'hypothèse d'indépendance, il y a un lien a étudier entre les CSP et HEB.

Table 11: Tableau des distribution conditionnelle 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 avec les profils lignes. On apprend par exemple dans ce tableau que 37.42% des ouvriers choisisent le camping comme herbergement de vacance.

Table 12: Tableau des distribution conditionnelle 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 parmis 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 13: 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

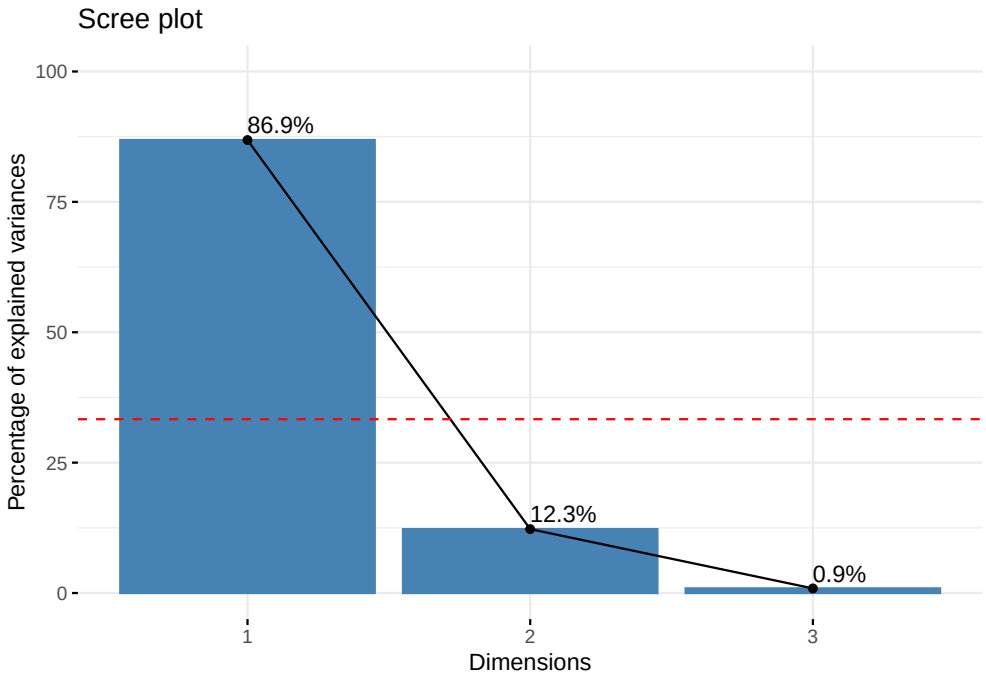


Figure 1: Visualisation des valeurs propres

Voici le tableau et le graphique de nos valeur propre. Quand on fait la somme des valeurs propres multiplié par n, on retrouve bien la statistique du khi2.

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 commençons avec les modalités de CSP.

Table 14: 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 nous montre les coordonnées que prendront les modalités de CSP sur les graphiques pour chaque dimension.

Table 15: 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

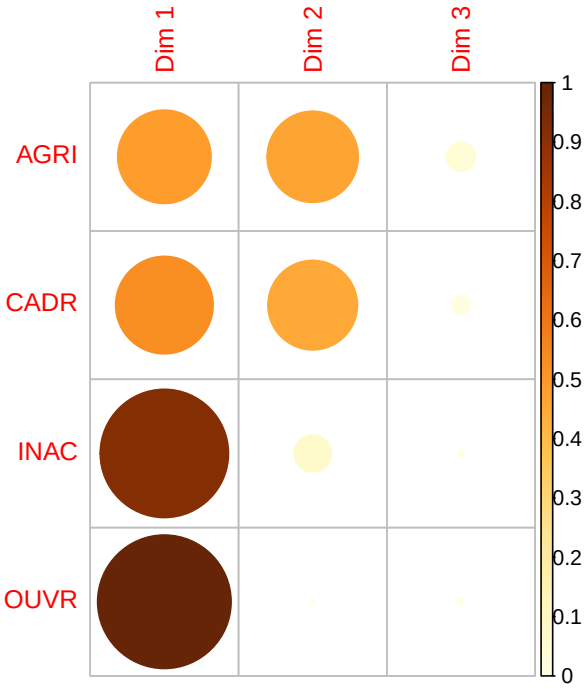


Figure 2: Visualisation des cos2 des modalités de CSP

On s'interesse 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 16: 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

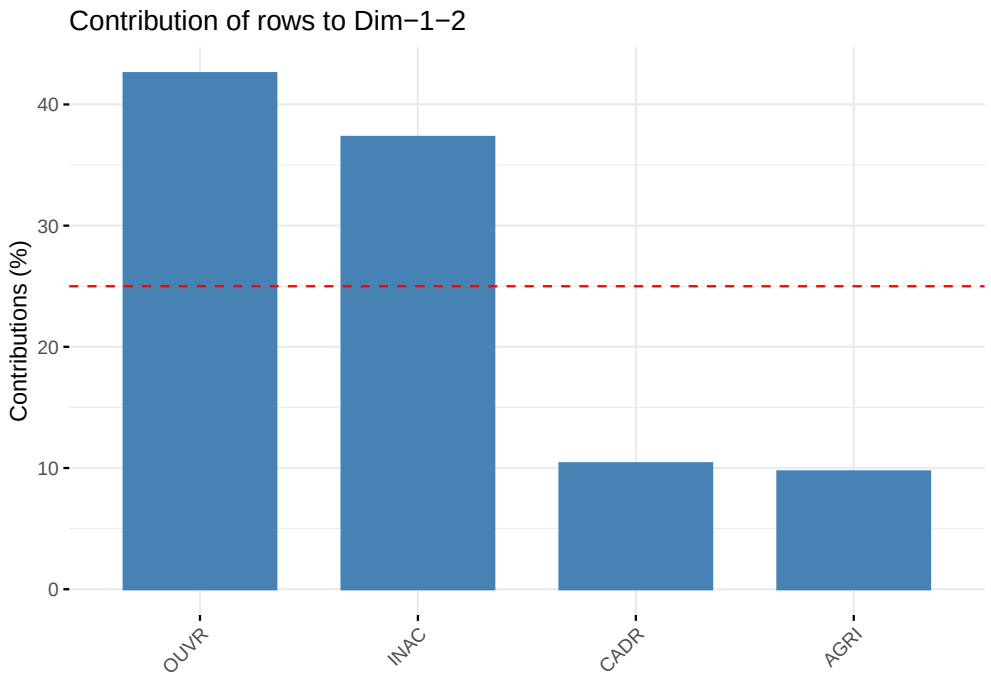


Figure 3: Visualisation des contributions des modalités de CSP

Regardons maintenant les comtributions. Le tableau nous montre le poucentage de la contribution pour chaque axe. Avec le graphique, on peut voir que les modalité OUVR et INAC sont les plus importantes dans la définition de le premier plan. La ligne poitillée rouge correspond à la combtribution moyenne.

Maintenant passons aux modalités de HEB.

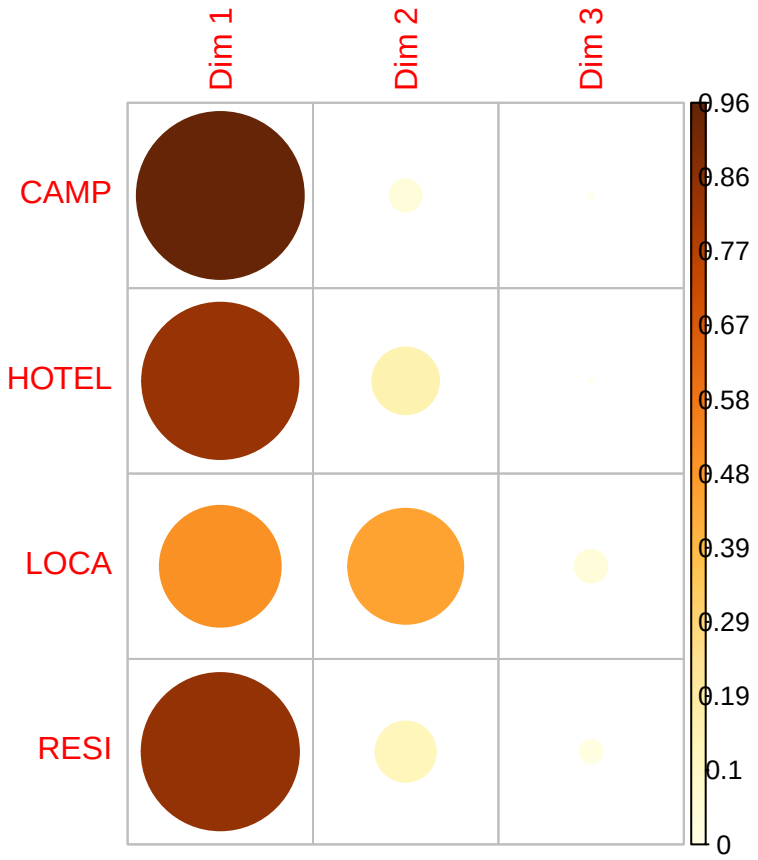
Table 17: Coordonnées des modalités de HEB

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

Table 18: 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



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

Table 19: 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

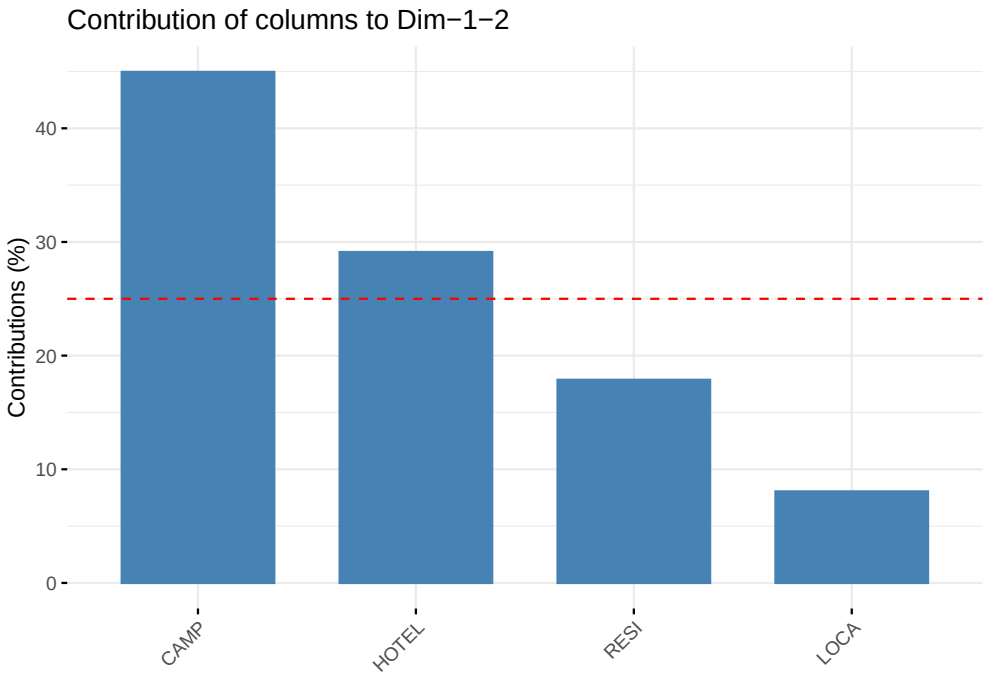


Figure 4: Visualisation des contributions des modalités de HEB

Regardons maintenant les comtributions. Le tableau nous montre le poucentage 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 de le premier plan. La ligne poitillée rouge correspond à la combtribution moyenne.

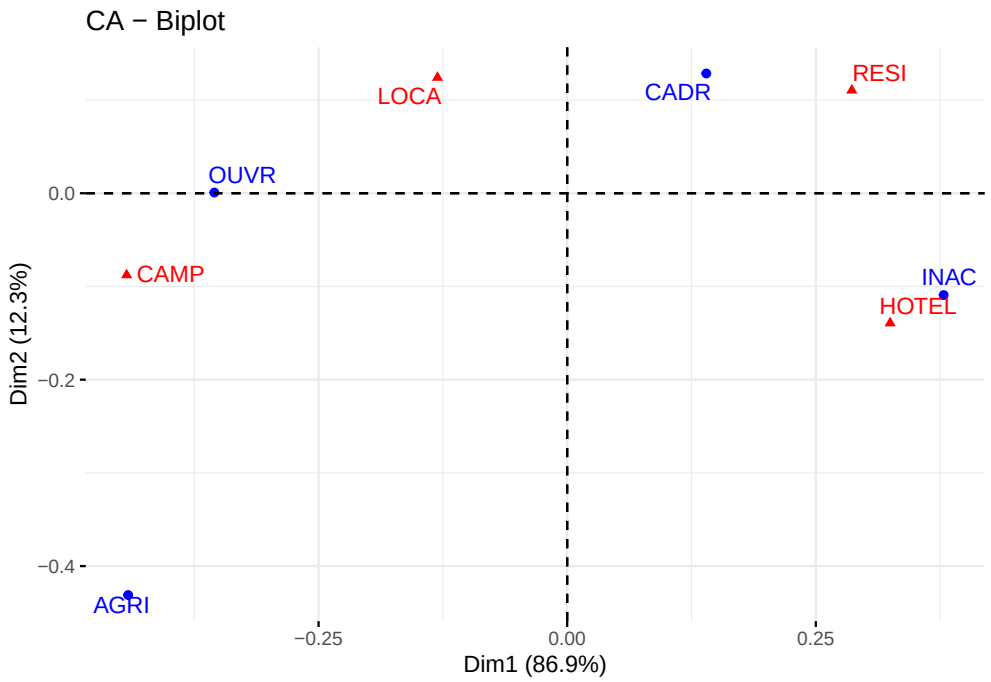


Figure 5: Bitplot

On peut ensuit 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'interesse uniquement aux modalité de CSP, on remarque les cadres et agriculteurs sont oposées, ce qui indique que leur profils s'oppose également.

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

La forme gernerale 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 groupe, on voit que les inactif sont liée au hotel, les cadres sont plus proche des résidences secondaire, et ouvrier opterons plus pour des campings. On peut imaginer qu'il y a un lien avec le cout de ces type 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 salarié (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 20: 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 21: 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 22: 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 23: Tableau des effectif théoriques

	none	light	medium	heavy
SM	0.2806606	-0.3526511	-0.2839006	0.4818124
JM	-0.7081704	-0.5842394	0.5063573	1.0926246
SE	2.2119849	-0.5484321	-1.0829559	-1.0139913
JE	-1.8607802	0.7686755	0.8897233	0.4742076
SC	0.7465200	0.0708205	-0.3638384	-0.6881439

Et aussi calculer la tableau des effectifs théoriques, utile pour le test du khi2.

Table 24: 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 25: 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 ligne et colonnes. Avec les profils lignes on voit que 28% des secretaries fume “moyennement”. Avec les profils colonnes, on remarque que parmis ceux qui ne fume pas, 40.98% sont Senior Employees.

On réalise ensuite l'AFC sur nos données.

Table 26: 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

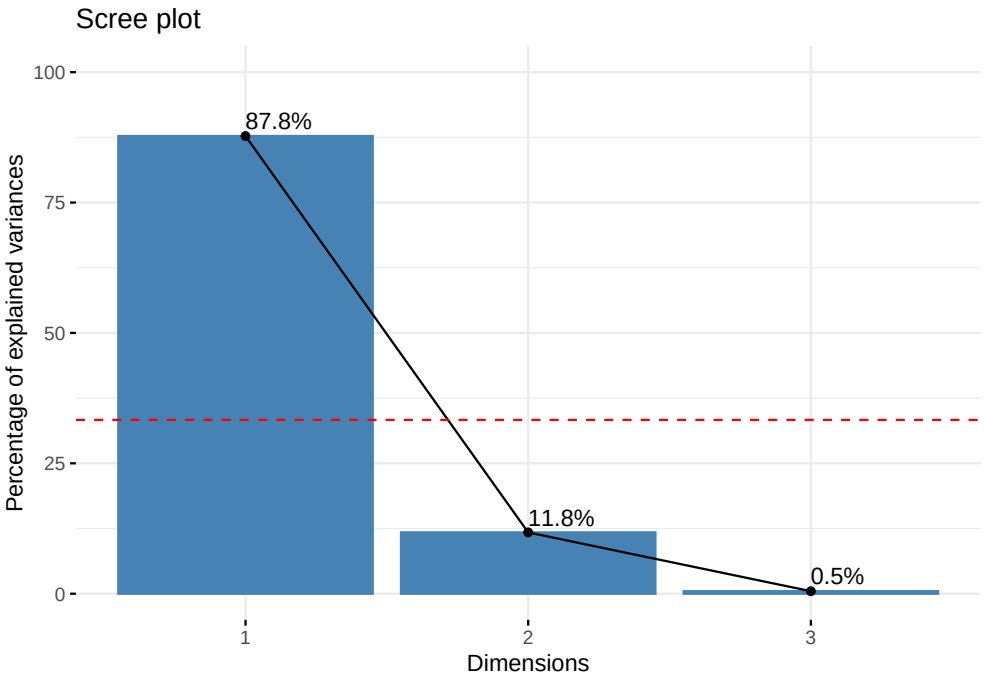


Figure 6: Visualisation des valeurs propres

On commence avec les valeurs propres. On voit dans le tableau et le graphique 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 27: 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 28: 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 les catégories de salarié, puis pour les catégories de fumeur.

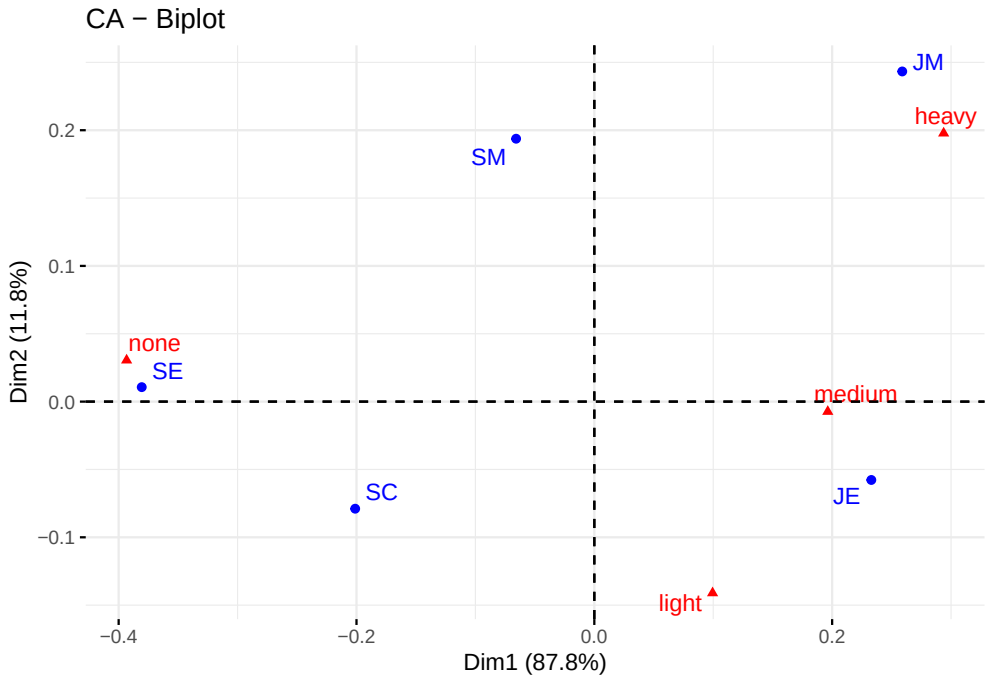


Figure 7: bitplot

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

On remarque qu'il n y pas de groupe qui se forme entre entre catégorie d'une meme variables, par contre pour les catégories de salarié on voit qu'il y a une opposition entre les Secretaries et les Junior Managers, ce qui s'ignifie que leurs profils s'oppose 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 "moyenement" et les Junior Employees, et entre les gros fumeurs et les junior Managers. Nous avons réussi à bien ciblé les lien qui existe entre la catégorie de salarié et la catégorie de fumeur.

Exercice 34

Il s'agit ici de proposer une méthodologie d'analyse textuelle pour identifier les auteurs de deux fragements de texte anonymes. On connaît pour chacun de ces fragments de texte la fréquence d'apparition de certaines lettres. On 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 spécifier.

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

On commence par voir si 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 29: 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

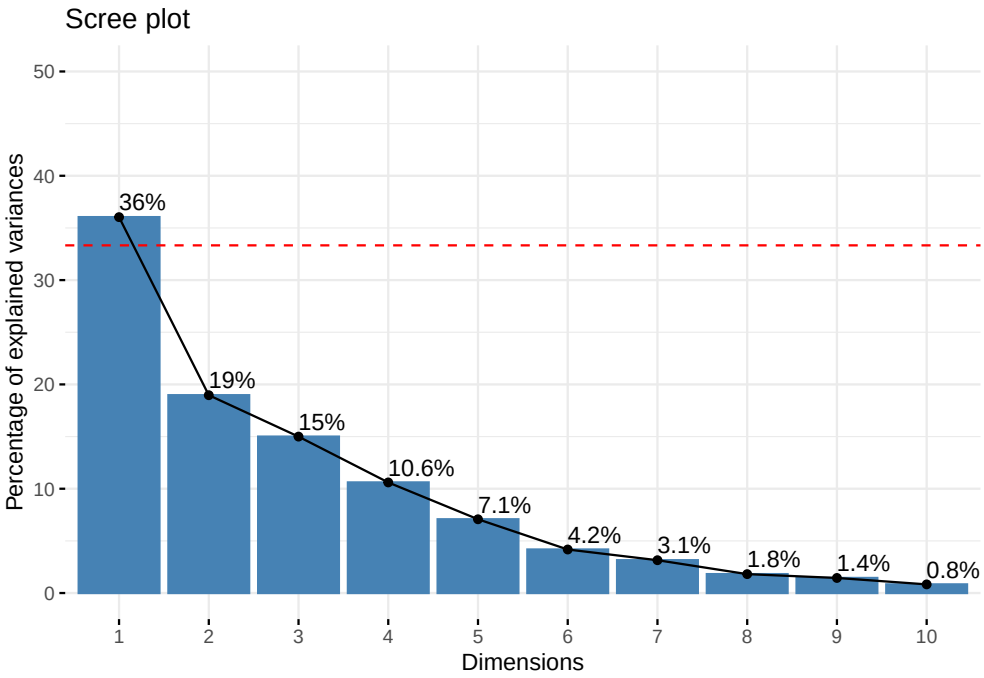


Figure 8: bitplot

On commence par determiner le nombre d'axe. 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 similaire quand on ajoute les individus supplémentaire, on conservera 4 axes égalements.

Table 30: 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 31: 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

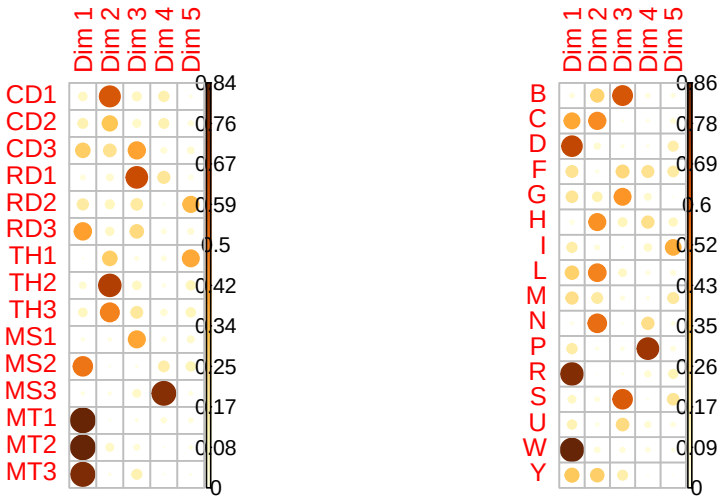


Figure 9: Visualisation des cos2 des auteurs et des lettres

On s'interesse maintenant à la qualité de représentation (cos2). On remarque avec le graphique et le tableau que les texte de Mark Twain on la meilleur représentations 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 32: 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 33: 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

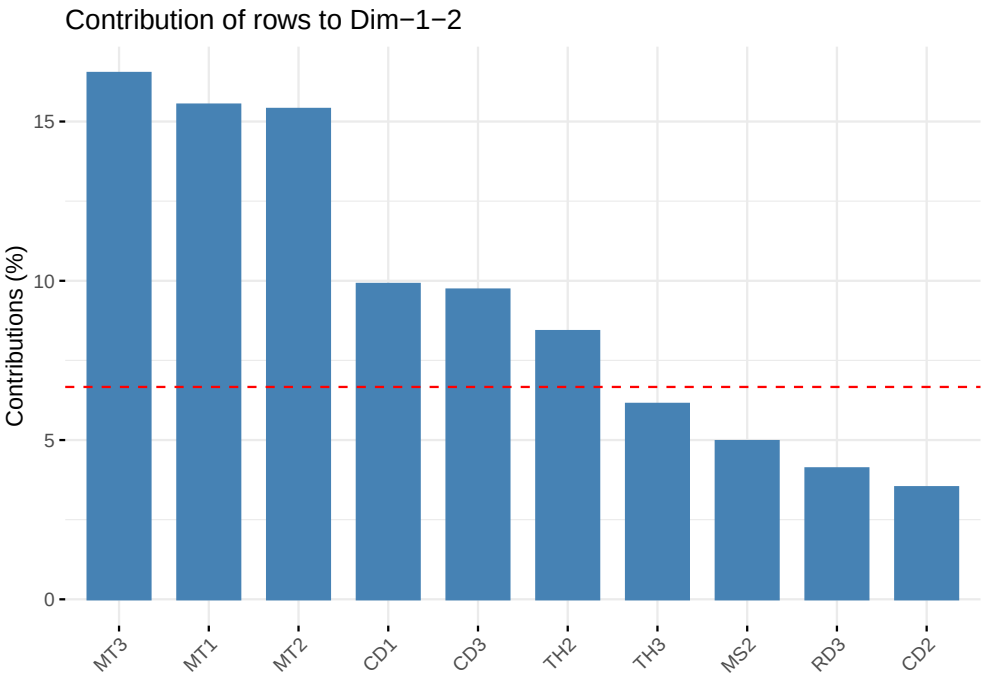


Figure 10: Visualisation des contributions des auteurs

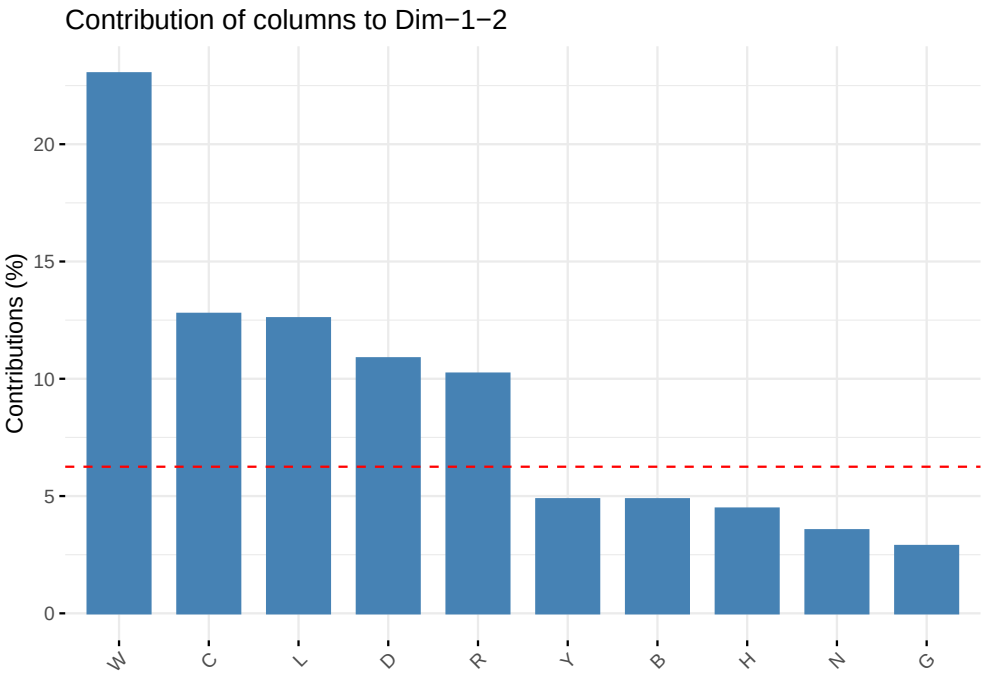


Figure 11: Visualisation des contributions des lettres

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 lettre W,C,L,D,R qui contribuent le plus au premier plan.

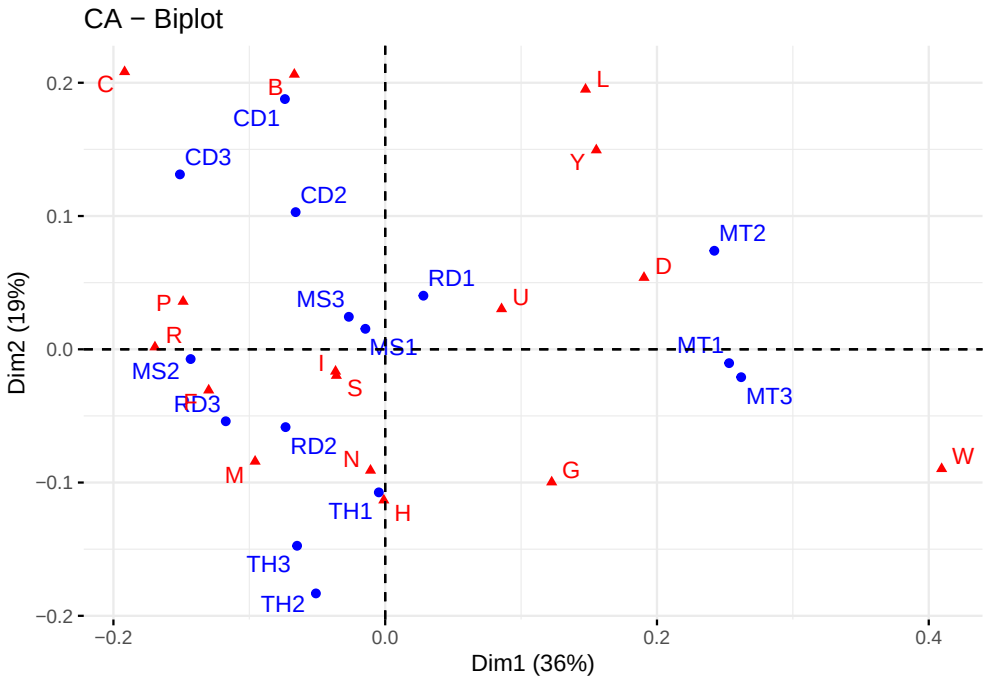


Figure 12: Bitplot

Quand on trace le bitplot, on remarque qu'il y a des groupes d'auteur qui se forme, associer à certaine lettres. Par exemple pour Charles Darwin, on voit que les lettres B et C sont celle ou il y a le plus de lien. Les auteurs René Descartes et Mary Shelley sont très lier, il semble avoir la meme utilisations 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 prononcé, il se confond presque avec René Descartes et Mary Shelley.

On réalise ensuite l'AFC avec les textes supplémentaire. On trouve des contributions et des qualité de représentation très similaires aux résultats précedents.

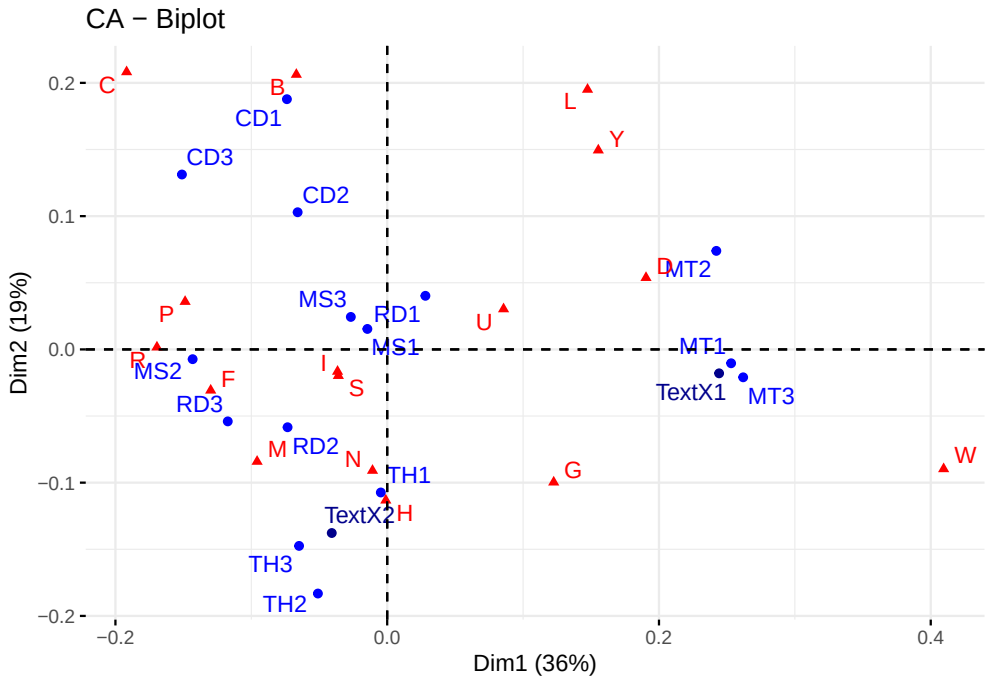


Figure 13: Bitplot avec les textes supplémentaire

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 alons classifier nos texte pour voir si cela se confirme.

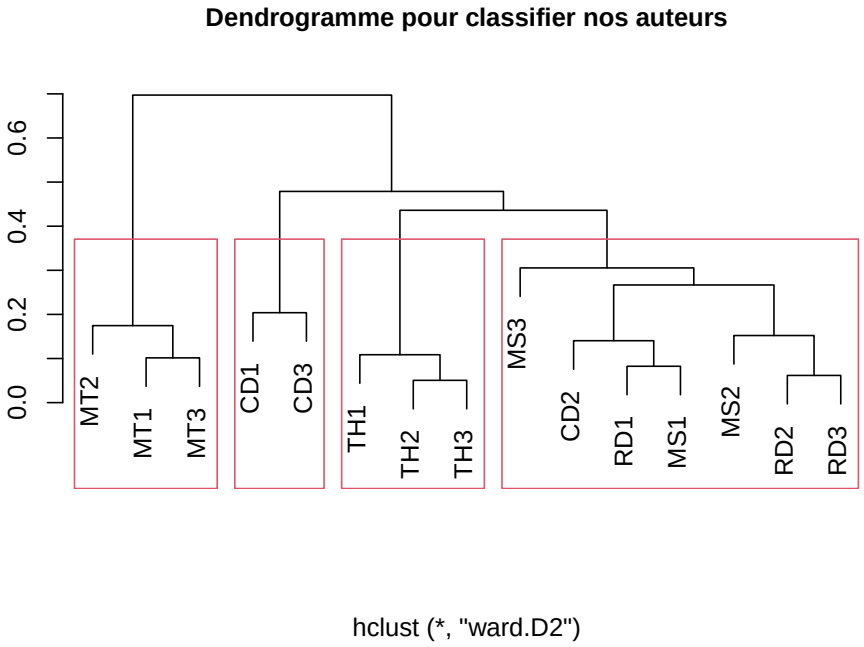


Figure 14: Dendrogramme pour classifier nos auteur

Le dendrogramme suivant nous montre la partition en 4 classes que nous offre nos données. Cela nous permet de confirmer les conclusions que nous avons tiré avec les bitplots. On voit qu'une classe est conposé des textes de t Mark Twain avec le texte 1 qui doit aussi être un texte de cet auteur. La deuxième classe comprte les textes de Thomas Hobbes avec le textes 2 qui doit être issu de cet auteur. La troisième classes est composée de deux textes de Charles Darwin, notre classification n'as pas considèrer le troisième texte de Charles Darwin dans cet classe. En effet ce texte ce 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.