Rapport d'étude

Etude du jeu de données MovieLens

Analyse descriptive du comportement des groupes d'utilisateurs

L3
CMI SID

Max Halford
Giovanni Zanitti

Professeur encadrant
Jonathan Louëdec

Table des matières

Table des figures

Chapitre 1

Introduction

Le but de l'analyse est de répondre à la question "**Qui aime quoi ?**". Le résultat sera une liste d'affirmations, par exemple, "*Les hommes qui ont la trentaine aiment les films d'actions*".

Il est important de remarquer qu'une affirmation peut-être :

— Plus ou moins vraie
— Plus ou moins détaillée

Evidemment seulement les affirmations qui sont nettement visibles seront gardées. Il n'est pas dit que des groupes d'utilisateurs existent. Au contraire le fait qu'il n'y en ait pas est aussi une information.

L'intuition devrait nous dire qu'on peut trouver des groupes plus ou moins précis. Le but est de trouver des groupes généralistes, peut-être dans l'optique de faire de la recommandation. Par exemple est-il plus intéressant de savoir que 70 pourcent des hommes du Midwest aiment les films d'actions que de savoir que 90 pourcent des avocats hommes du Midwest ayant la trentaine aiment les films d'action-aventure ? Sachant que le deuxième groupe est plus restreint que le premier.

Les données appartiennent à deux entités, les films et les utilisateurs. L'objectif est d'étudier les liens entre ces entités. L'étude se fait en deux parties. Premièrement on construit des Analyses Factorielles des Correspondances (AFC) pour pouvoir analyser des liens entre les variables de chaque entité. Les films ont deux variables : leur genre et leur date de sortie. Les utilisateurs ont trois variables : l'âge, l'occupation professionnelle et la région d'habitation. Il y'a donc 6 (2×3) AFCs possibles. Cependant nous avons aussi pour faire chaque AFC deux fois, une fois pour chaque sexe. En tout on a donc 12

(6×2) AFCs à analyser. Deuxièmement on construit une Analyse des Correspondances Multiples (ACM) pour observer des liens entre plus de deux variables. Avec l'ACM on pourra aussi classifier les individus par groupe en construisant un dendogramme. De cette façon on aura une vision globale grâce aux AFCs et des informations plus précises grâce à l'ACM.

Chapitre 2

Choix des données et munging

Tout d'abord **seulement les notes de 4 et 5 sont conservées**. En effet le but étant de connaitre les préférences des utilisateurs, il nous paraît pertinent de se limiter à l'étude des bonnes notes.

Ensuite il s'agit de choisir les variables à étudier. Pour les utilisateurs sont gardés :

— L'âge
— Le sexe
— L'occupation professionnelle
— Le code ZIP

Le munging est comme suit : l'âge devient une variable qualitative en ne conservant que le premier chiffre de l'âge (35 devient 30, 42 devient 40 etc.). Puisque le sexe est une variable binaire, les utilisateurs sont scindés en deux groupes dans la construction des AFCs, en effet cela permet de rajouter une variable aux AFC qui étudient les relations entre deux variables. L'occupation professionnelle reste intacte. Chaque code ZIP est associé à l'état américain qui lui est propre ou bien à la valeur *Other* pour les codes éronnés et les pays étrangers aux Etats-Unis.

Pour les films sont gardés :

— La date de sortie
— Le(s) genre(s)

Pour le munging la décennie de sortie du film est extraite. Un film peut avoir plusieurs genres et donc un choix s'impose. Pour les AFCs les tableaux de contingence prennent en compte le fait qu'aimer un film qui a deux genres

équivaut à aimer un film du premier genre et un autre film d'un autre genre. Pour l'ACM on procède de la même façon.

Tout le munging est effectué en *Python* avec l'aide du module *pandas*. Une jointure est effectuée sur les trois fichiers csv fournis. Des tableaux de contingence pour les AFCs sont sauvegardés sous format csv, les analyses sont effectués avec *R*. Les données sont filtrées comme indiqué précédemment. Les mêmes tableaux de contingence sont construits pour les hommes et les femmes. L'API *api.zippopotam.us* est utilisée pour convetir les codes ZIP en états américains qui sont ensuite filtrés selon la carte suivante.

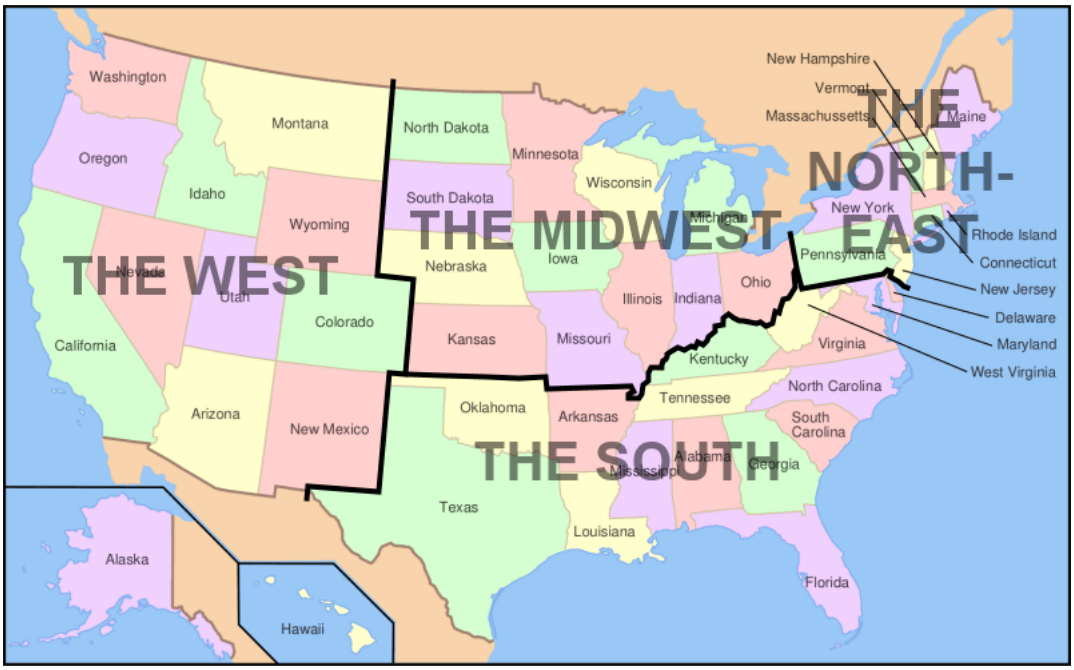

FIGURE 2.1 – Régions des Etats-Unis

Toutes les données sont sauvegardées comme suit :

— Un fichier csv contenant toutes les données.
— Une fichier csv contenant les données utiles à l'ACM.
— Une base de donnée MongoDB contenant toutes les données.

Le premier fichier sert à construire les tableaux de contingence pour produire les AFCs. Le deuxième diffère du premier car il ne contient pas toutes les données, de plus les genres de films sont séparés (par exemple un film ayant trois genres devient trois lignes). La base MongoDB est utilisée pour deux

raisons. D'une part la création des tableaux de contingence pour les AFCs relatifs aux genres des films est très rapide, en effet le module *pymongo* renvoit un curseur Python, un type de données natif qui permet d'itérer plus rapidement que sur un tableau de données. D'autre part une base de données peut servir dans le cas de l'étude d'une sous-population particulière d'une grande population, en effet les temps de calculs pourraient être long si on élargissait notre étude au jeu de données de 10 millions de lignes de MovieLens. De la même façon calculer l'ACM avec R pourra se révéler hardu sur le jeu de données de 10 millions de ligne, pour cela il serait envisageable d'utiliser *Julia*, notamment grâce à https ://github.com/MaxHalford/MultivariateStats.jl.

Chapitre 3

Analyses Factorielles des Correspondances

Pour rappel nous avons construits 12 AFC. Nous avons décidé de faire une AFC par sexe pour ensuite pouvoir les comparer et voir si des groupes d'individus sont stables selon cette variable. Les AFCs vont présenter des liens entre trois variables, dans le but de nous faire une idée de profils-types des personnes aimant des films.En regardant les contributions de chacunes des variables sur les axes, nous avons pu trouver des relations entre elles. Le tableau suivant récapitule les résultats obtenus :

AFC	Contribution du plan	Axe 1	Axe 2
M_genreVSage	88,1%	Oppose les jeunes aimant des films d'actions aux plus anciens qui regardent plutôt des films noirs.	Oppose les hommes de 40 ans aimant des comédies aux cinquantenaires aimant des thrillers, films d'horreurs et crimes.
F_genreVSage	91,6%	Oppose les enfants filles aimant des films d'enfants aux femmes de 50 ans aimant des films dramatiques.	Oppose les femmes de 40 ans aimant des films d'action à celles de 50 ans aimant des comédies.
M_genreVSoccupation	66,2%	Oppose les étudiants aimant des films d'actions aux libraires et educateurs aimant des films dramatiques.	Oppose les ingénieurs aimant des films d'enfant contre les écrivains aimant des thrillers, films d'horreurs et crimes.
F_genreVSoccupation	67,8%	Oppose les étudiantes aimant les films d'actions et d'horreur aux educatrices et libraires aimant les films dramatiques.	Oppose les femmes programmeurs aimant les films d'enfants et les comédies aux étudiantes préférant les films d'horreurs et films noirs.
M_genreVSregion	87,9%	Oppose hommes du Midwest aimant des films d'actions aux les hommes habitant au Nord-Est aimant les films dramatiques.	Oppose les hommes de l'Ouest aimant des films familliales (Comédies, musicales, animations) aux autres aimant plutôt des thrillers, films d'horreurs et crimes.
F_genreVSregion	89,4%	Oppose les femmes du Nord-Est aimant des drames aux femmes du sud aimant plutôt des films d'actions et thrillers.	Oppose les femmes de l'ouest aimant des films pour enfant aux autres femmes aimant des films d'horreurs.
M_releaseVSage	98,0%	Oppose les jeunes aimant des films des années 90 aux plus vieux qui aiment des films des années 50.	Oppose les enfants aimant des films des années 90 aux trentenaires aimant des films des années 80
F_releaseVSage	94,8%	Oppose les enfants aimant des films des années 90 aux femmes de 40 ans aimant des films des années 50.	Oppose les femmes de 40 ans aimant des films des années 60 aux jeunes femmes de 20 ans aimant des films des années 80.
M_releaseVSoccupation	93,5%	Oppose les étudiants aimant des films des années 90 aux educateurs aimant les films des années 50.	Oppose les ingénieurs et programmeurs aimant les films des années 80 aux autres regardant des films des années 50.
F_releaseVSoccupation	85,8%	Oppose les étudiantes et femmes aux foyers regardant des films des années 90 aux educatrices aimant des films des années 50.	Oppose les femmes d'affaires regardant des films des années 80 aux écrivaines regardant des films des années 50.
M_releaseVSregion	98,8%	Oppose les hommes du sud aimant des films des années 90 aux hommes du nord-est aimant des films des années 50.	Oppose les hommes du Midwest aimant les films des années 80 aux hommes de l'ouest aimant les films des années 50.
F_releaseVSregion	92,3%	Oppose les femmes du sud aimant les films des années 90 aux femmes de l'ouest aimant des films des années 80 et 50.	Oppose les femmes du Midwest aimant les films des années 80 aux femmes de l'ouest aimant les films des années 30.

FIGURE 3.1 – Commentaires sur les AFC

Lorsqu'on regarde les AFC une à une, on peut assez facilement les commenter en se basant sur la contribution des variables par rapport aux axes du plan. De plus, ces analyses sont significatives puisqu'on peut voir que leurs contribution, autrement dit la part de variance expliqué, est très souvent au dessus de 80%. Cependant, dès lors qu'on essai de comparer ces AFC les unes avec les autres, les conclusions sont plus difficiles à tirer.

En effet, La lecture de ce tableau n'est pas très facile dû à la quantité d'information fourni. Pour chaque AFC, on constate deux oppositions entre deux catégories de personnes. Ce qui, avec 12 AFC réalisées, ne nous donne pas de profils-types des individus ayant donné une note de 4 ou 5 mais plutôt nous amène à mélanger trop d'informations et par conséquent, tirer des conclusions peu claires et même éronées. C'est pourquoi il est indispensable d'essayer de voir des similarités entre ces AFC, afin d'avoir un commentaire efficace sur l'ensemble des données.

Par exemple, on peut voir que ce sont souvent les mêmes genre de films qui reviennent pour la descrpitions des axes (action, drama, children et thriller). Cela nous montre que ce sont les genres qui influent le plus sur la bonne note de l'utilisateur. On voit aussi que les Thrillers, films d'horreurs et les films de type crime sont associés à 2 reprises (M_genreVSage et M_genreVSoccupation). C'est une observation qui parait plutôt logique car ce sont des genres qui se ressemblent.

On peut aussi clairement voir que l'age de l'utilisateur joue un rôle important quand à la date de réalisation d'un film qu'il aime. En effet, plusieurs fois nous observons que les adolescents, les jeunes (personnes de 20 ans) et de fait les étudiants aiment des films des années 1990, quelque soit le sexe.

Concernant les différences entre les sexes justement, elle est difficile de voir une ressemblance flagrante sur toutes ces AFC. Appart peut être sur l'AFC releaseVSoccupation où les oppositions par rapport aux axes sont quasiment identique quelque soit le sexe. Ces résultats nous laisse penser que le sexe est une variable explicative sur la bonne note de l'utilisateur puisqu'elle influe sur les autres associations de variables sauf releaseVSoccupation.

Enfin, ces résultats nous permettent de faire des groupes de personnes de manière significative avec deux variables. Cependant, le but de notre étude est de trouver des classes plus élaborées avec des critères plus globaux.

Afin d'avoir un discours plus clair sur ces données, nous avons utilisé une autre technique statistique plus générale : l'Analyse à Composante Multiple (ACM). Celle-ci nous permettera de confronter plus de 2 variables sur un même graphique contrairement à l'AFC. Les analyses que nous pourrons faire ressembleront à celles faite à propos de ces dernières puisque l'AFC est un cas particulier de l'ACM.

Chapitre 4

Analyse des Correspondances Multiples

4.1 Construction de l'ACM

L'ACM est une généralisation de l'AFC. Elle a pour but d'étudier la distance entre des individus au travers de la distance du χ^2. En effet l'ACM est une ACP avec une différente métrique. Comme l'Analyse en Composantes Principales elle étudie les axes qui maximisent la variances des données.Pour s'épargner des complications et pour gagner du temps on a décidé d'utiliser la librairie *FactoMineR*. On a aussi utilisé la librairie *ggplot2* pour rendre les graphiques plus agréables à regarder. Avec ces deux outils le code n'est pas trop verbeux. Il est disponible à ce lien `https://github.com/MaxHalford/MovieLens/blob/master/MCA.R`.

4.2 Observations initiales

L'ACM fournit plusieurs informations. On peut commencer par regarder les contributions à la variance des axes.

Axe	Contribution
1	3.1798946
2	2.9845177
3	2.6680496

Les axes principaux n'expliquent qu'à eux trois 8.832462% de la variance. On ne peut pas interpréter de façon graphique les données. On ne pourra pas

non plus interpréter les contributions individuelles des variables aux axes. La raison est que nous avons trop de variables. Une approche serait de faire des ACMS spécifiques sur certaines variables. Cependant, ayant déjà construits des AFCs, nous allons faire une classification hiérarchique pour conserver les informations de chaque axe et comprendre quelles variables divisent les individus.

4.3 Classification hiériarchique

Pour une analyse en composantes principales donnée, on peut regrouper les individus similaires au moyen d'une classification des données. La classification étudie les coordonnées des individus sur le nombre d'axes souhaités. Sous R, la fonction `HCPC` de *FactoMineR* est très lourde. On va donc reconstruire une ACM de 5000 votes tirés aléatoirement. On suppose que cet échantillon est représentatif de par son caractère aléatoire. On garde les 40 premiers axes de l'ACM pour représenter environ 85% de la variance totale. On pourrait garder tous les axes et alors l'ACM n'aurait servi à rien mais la classification décrirait *trop* bien les individus analysés et on ne pourrait pas généraliser nos conclusions à tous les individus. On obtient le dendogramme suivant.

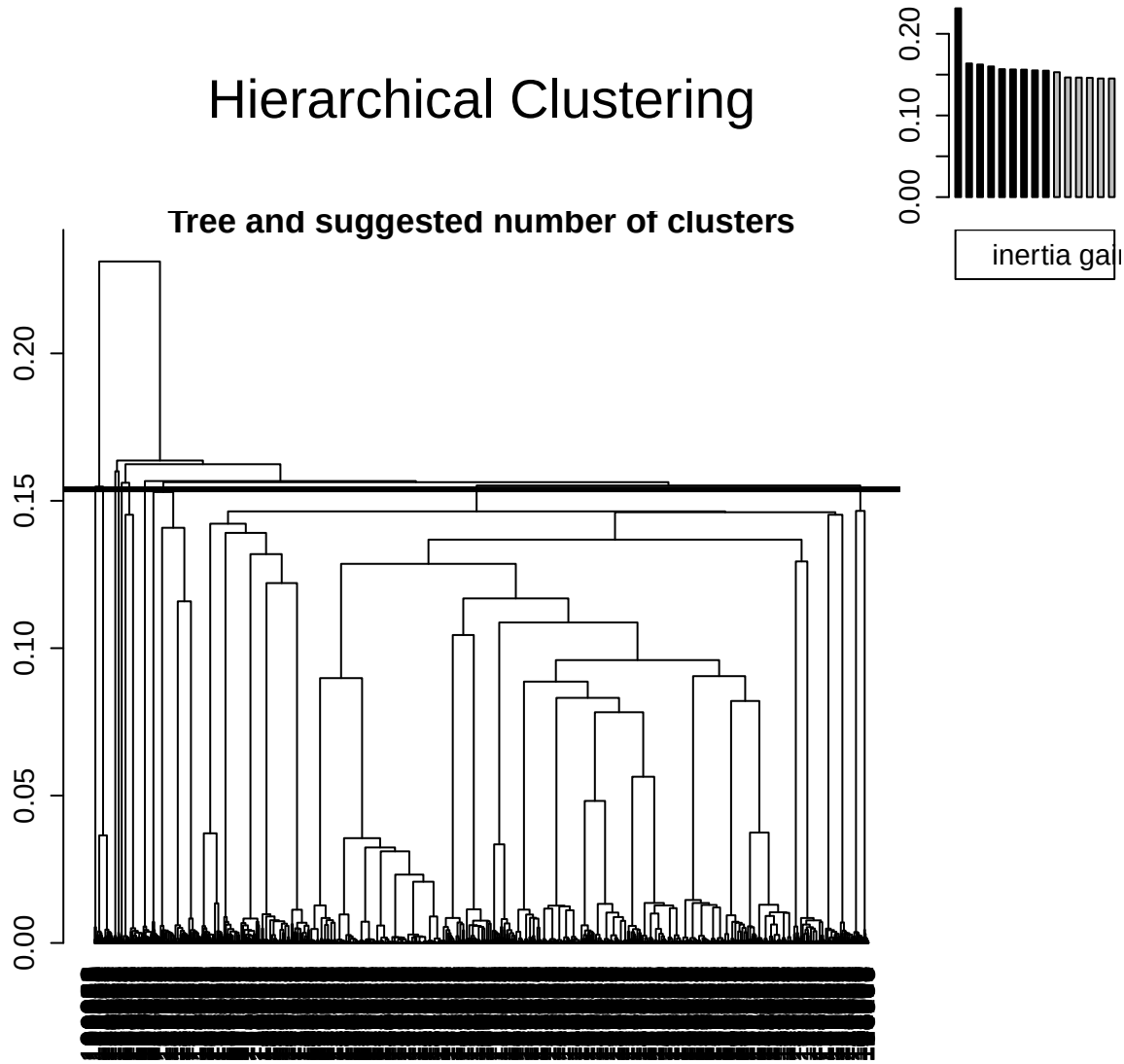

FIGURE 4.1 – Classification hiérarchique d'un échantillon de taille 5000

Le test du khi^2 sur les variables donne le tableau trié par p-value suivant :

Variable	p-value	Degrés de liberté
Occupation	$0.0e+00$	180
Catégorie d'âge	$0.0e+00$	54
Genre	$0.0e+00$	153
Région d'habitation	$1.125180e-24$	9
Sexe	$2.209744e-21$	36
Décennie de sortie	$3.960011e-11$	63

Les variables *Occupation*, *Catégorie d'âge* et *Genre* sont celles qui caractérisent le plus la séparation des individus. On constate que la décennie de sortie influe moins la séparation.

La classification sépare les individus selon dix groupes. Le huitième groupe est immense relativement aux autres. En effet il regroupe autour de 80% des utilisateurs(*trices*). Grâce à *FactoMineR* on peut obtenir les votes représentant le mieux chaque groupe, les *parangons*. Bien sûr nous avons construit le dendogramme avec un échantillon de 5000 votes, donc les parangons se ressembleront mais ne seront pas les mêmes. Nous avons refait la classification avec 5 échantillons et nous avons obtenu les mêmes classes. La classification proposée par *FactomineR* est donc robuste.

Groupe	Parangon	Variables
1	60139	M none Midwest 20s 1990s Drama
2	80251	M student West 30s 1990s Drama
3	41038	M student North-East 20s 1990s Documentary
4	25045	M doctor Midwest 30s 1990s Drama
5	21522	F homemaker West 30s 1990s Drama
6	60871	M technician West 30s 1990s Drama
7	58496	F student Midwest 20s 1990s Fantasy
8	75464	M artist Midwest 30s 1990s Drama
9	118859	F healthcare South 30s 1990s Drama
10	105256	M retired Midwest 50s 1990s Drama

Pour chaque classe on récupère les modalités caractéristiques, c'est à dire pour lesquels les p-values du test de Fisher sont inférieures à 0.05. La description des 10 classes est disponible sur les deux pages suivantes.

Classe 1

occupation = none, programmer, administrator, engineer, educator, other, student
ageCategory = 20s, 40s, 50s
genre = NULL
region = Midwest, West, North-East
gender = NULL
releaseDecade = 1990s

Classe 2

occupation = student, other, administrator, programmer, engineer, librarian,
educator, writer, artist, executive, lawyer, homemaker, doctor, salesman, none,
retired, marketing, healthcare, entertainment, scientist, technician
ageCategory = 10s, 20s, 30s, 60s, 70s
genre = Comedy, Documentary, Fantasy, Film-Noir, Romance, War
region = Midwest, West
gender = F, M
releaseDecade = 1920s

Classe 3

occupation = NULL
ageCategory = NULL
genre = Documentary, Drama
region = North-East
gender = NULL
releaseDecade = NULL

Classe 4

occupation = engineer, doctor, programmer, educator, other, student
ageCategory = 10s, 20s ,30s
genre = NULL
region = Midwest, North-East
gender = F, M
releaseDecade = NULL

Classe 5

occupation = homemaker, student
ageCategory = NULL
genre = Thriller
region = Other, Midwest, South
gender = F, M
releaseDecade = 1990s

FIGURE 4.2 – Classes hiérarchiques 1 à 5

Classe 6

occupation = entertainment, lawyer, retired, technician, scientist, marketing, healthcare, artist, executive, writer, librarian, administrator, programmer, educator, engineer, other, student
ageCategory = 10s, 30s, 50s, 60s
genre = Crime, Film-Noir
region = Other, South, Midwest
gender = F, M
releaseDecade = 1920s

Classe 7

occupation = NULL
ageCategory = 60s
genre = Fantasy, Comedy, Drama
region = West
gender = NULL
releaseDecade = 1980s

Classe 8

occupation = artist, marketing, retired, healthcare, entertainment, scientist, librarian, writer, technician, administrator, educator, other, engineer, programmer, student
ageCategory = 10s, 30s, 60s
genre = War, Adventure, Horror
region = North-East, South
gender = NULL
releaseDecade = 1940s, 1990s

Classe 9

occupation = healthcare, salesman, artist, technician, executive, writer, librarian, administrator, programmer, engineer, educator, other, student
ageCategory = 10s, 20s, 40s, 50s
genre = Musical
region = Other, Midwest, South
gender = F, M
releaseDecade = NULL

Classe 10

occupation = retired, educator, scientist, artist, technician, administrator, student
ageCategory = 10s, 20s, 30s, 40s, 60s, 70s
genre = Drama, Western Mystery
region = West, South, Midwest
gender = F, M
releaseDecade = 1990s

FIGURE 4.3 – Classes hiérarchiques 6 à 10

Les tailles des classes sont les suivantes.

Classe	Taille observée	Pourcentage d'importance
1	46	0.92
2	3987	79.74
3	22	0.44
4	37	0.74
5	17	0.34
6	355	7.1
7	25	0.5
8	247	4.94
9	115	2.3
10	149	2.98

On remarque que la deuxième classe est de loin la plus large. Cela confirme les observations des AFCs dans le sens ou il est difficile de classifier les individus car ils sont assez similaires. On peut tout de même noter qu'après avoir fait 5 classifications les classes ne changeait pas beaucoup, on peut donc supposer que les classes proposées partagent bien les votes.

Chapitre 5

Conclusion

Le but de cette étude était de voir s'il existait des groupes de personnes de même profil qui aime le même genre de film. C'est pourquoi, dès le début, nous avons filtrer nos observations sur les notes 4 et 5. Après avoir manipulé les données et utiliser différentes méthodes statistiques telles que l'AFC, l'ACM et la classification, notre conclusion est limité.

En effet, lorsque nous avons regarder les résultats des AFCs, nous avons vu que nous pouvions réaliser des groupes de personnes suivant 2 critères. Ces groupes sont d'ailleurs significatif à la vue des différentes contributions des axes pour chacune de ces AFC. Cependant, notre objectif était de trouver des groupes plus généraux, c'est pourquoi nous avons réaliser une ACM.

Au regard des résultats de cette dernière, nous n'avons pas pu interpréter de façon graphique les données puisque les 3 premiers axes expliquent seulement 8.8% de la variance. La raison est que nous avons trop de variable pour que l'ACM Soit interprétable sur un si petit nombre d'axe.

Enfin, nous avons quand même réalisé une classification hiérarchique en prenant les 40 premiers axes de l'ACM qui représentent au cumulé 85% de la variance totale. De plus, afin de réduire les temps de calcul, nous avons limité le nombre d'observations à 5000 votes tirés aléatoirement que nous avons supposé représentatif. La conclusion de la classification est que les individus sont très similaires, mais que les classes en marge se démarquent bien. Il serait à priori difficile d'effectuer du machine learning sur ces données puisqu'elles sont très regroupées et difficilement séparables.