UNIVERSITÉ
DE REIMS
CHAMPAGNE-ARDENNE

Analyse des Préférences du Public et des Caractéristiques des Films sur IMDb en 2018

ERTAS Elif

Master de Statistique pour l'évaluation et la prévision

TABLES DES MATIERES

PARTIE A : Compte rendu de l'analyse 3
I. Introduction 3
II. Description des données 3
III. Importation et préparation des données 5
IV. Analyse de données 6
V. Conclusion 8
PARTIE B : Annexes 9
- Annexe 1 : Lien vers l'open data IMDb 9
- Annexe 2 : Modèle Conceptuel de Données 9
- Annexe 3 : Dictionnaire des variables 10
- Annexe 4 : Répartition des groupes 10
- Annexe 5 : Résumé statistique 11
- Annexe 6 : Genres les plus fréquents 11
- Annexe 7 : Catégorie professionnelles les plus fréquents 12
- Annexe 8 : Variance expliquée par chaque dimension dans l'AFM 12
- Annexe 9 : Contribution des groupes aux dimensions dans l'AFM 13
- Annexe 10 : Visualisation de la contribution des groupes 13
- Annexe 11 : Visualisation des axes partiels 14
- Annexe 12 : Cercle des corrélations de l'AFM 15
- Annexe 13 : Graphique des individus de l'AFM 16

PARTIE C : Codes 17
➔ R Studio (Préparation de la base finale) 17
➔ R Studio (analyse) 20

PARTIE A : Compte rendu de l'analyse

I. Introduction

IMDb, acronyme pour **Internet Movie Database**, est une plateforme en ligne incontournable dans l'industrie cinématographique. Lancée en 1990, cette base de données exhaustive recense une vaste collection d'informations sur des films, des émissions de télévision, des séries, des acteurs, des réalisateurs et bien plus encore. Son objectif principal est de fournir aux utilisateurs une source fiable et détaillée sur le monde du divertissement.

Actuellement, IMDb est largement utilisé au quotidien par des millions de passionnés du cinéma et de la télévision à travers le monde. Cette plateforme constitue une référence majeure, offrant des critiques, des évaluations, des synopsis, des informations sur les équipes de production et des données diverses pour les amateurs de culture audiovisuelle.

Une particularité d'IMDb est sa mise à disposition des données ouvertes (Annexe 1) pour le grand public, permettant ainsi aux chercheurs, analystes et passionnés de cinéma d'accéder à un ensemble de données considérable pour des études ou des analyses.

Dans le cadre de cette étude, nous avons délibérément choisi de nous concentrer sur les films sortis exclusivement en 2018. Ce choix découle de la volonté de minimiser les biais potentiels liés à la période de la pandémie de Covid-19, où les sorties de films étaient moins fréquentes et les dynamiques du marché étaient perturbées. En outre, cela permet d'éviter de considérer les sorties plus récentes, alors que le monde culturel se rétablit progressivement.

II. Description des données

Un travail préliminaire est nécessaire pour établir une base solide. En effet, nous avons choisi de combiner quatre bases de données distinctes (Annexe 2) afin d'agréger toutes les informations nécessaires pour chaque titre.

La base de données finale est composée de 1292 titres définis par 13 variables (Annexe 3). Nous avons sélectionné des variables clés pour notre étude. Parmi celles-ci, le titre du film, ses genres, sa durée, les réalisateurs et les évaluations des utilisateurs ont été des points focaux. Ces variables ont été analysées pour comprendre l'impact du genre, de la durée et de la direction sur la perception et la performance des films.

Ce graphique met en lumière que près de 50% des données répertoriées sont des films, suivis par d'autres catégories telles que les courts-métrages, les vidéos, les séries, etc.

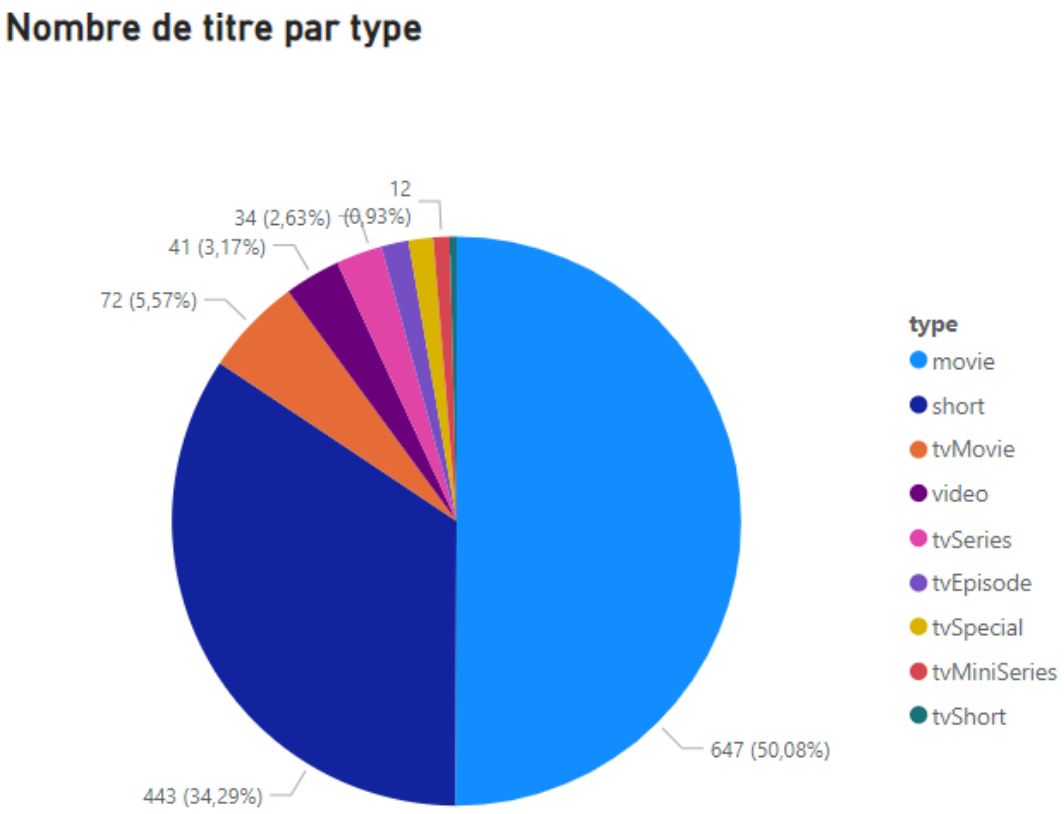


Figure 1 : Graphique camembert qui illustre la répartition des titres par type sur la plateforme IMDb pour l'année 2018

En complément de la répartition des types de titres sur IMDb pour l'année 2018, un deuxième graphique met en évidence le nombre de votants par type de titre. Ce graphique révèle une tendance significative : les films ont recueilli une part prépondérante des votes par rapport aux autres catégories de contenus cinématographiques.

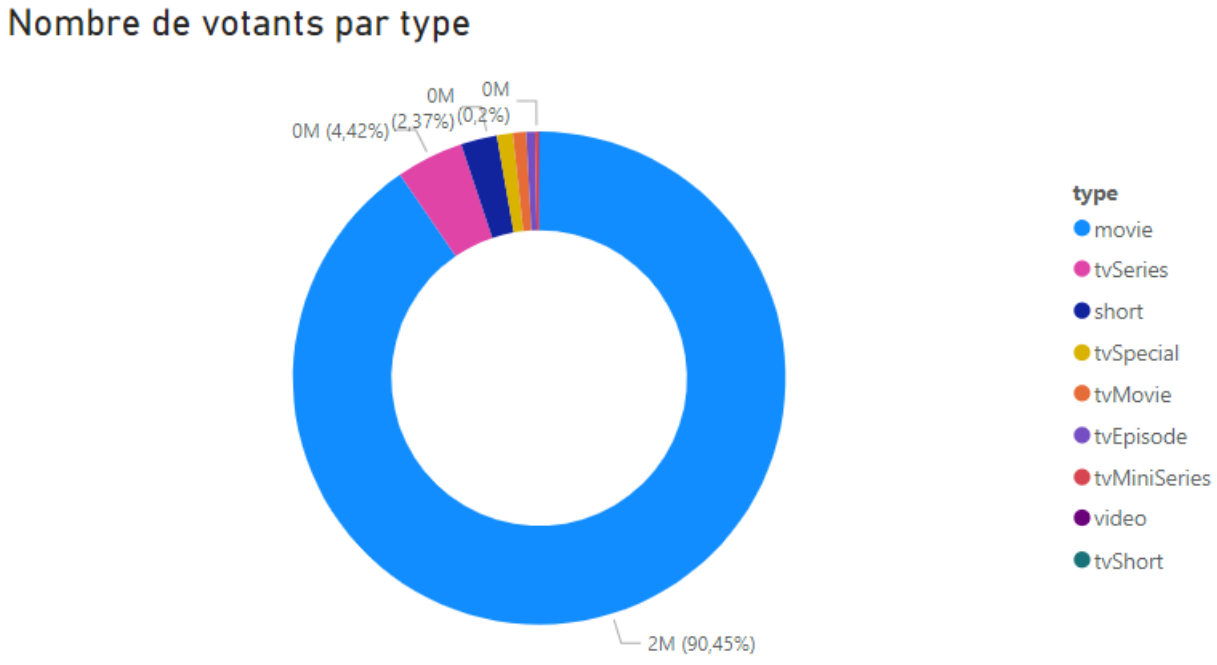


Figure 2 : Graphique camembert qui illustre le nombre de votants par type sur la plateforme IMDb pour l'année 2018

Ces observations ont motivé notre décision de concentrer notre analyse uniquement sur les films. En effet, cette prédominance significative des films dans notre base de données initiale suggère qu'ils constituent une part substantielle des productions répertoriées pour cette année spécifique.

III. Importation et préparation des données

Nos données sont enregistrées dans un fichier CSV où la virgule est utilisée comme délimiteur. Cependant, la variable 'deathYear', qui indique l'année de décès des réalisateurs associés à chaque titre, présente un nombre considérable de valeurs manquantes (97,6 %). Ce constat suggère que la grande majorité des réalisateurs répertoriés dans notre base sont vraisemblablement encore en vie. Par conséquent, ayant relevé un manque de pertinence pour notre analyse, nous avons pris la décision de retirer cette variable de notre ensemble de données.

En outre, dans notre démarche d'analyse, nous avons choisi de restreindre notre étude aux films, ce qui a abouti à une sélection de 647 titres pour notre analyse ultérieure. Cette sélection nous a permis de concentrer notre attention sur cette catégorie spécifique de contenu cinématographique, offrant ainsi une perspective plus ciblée et détaillée pour notre étude.

Pour optimiser notre analyse, nous avons procédé à une transformation des variables 'genres' et 'primaryProfession'. Étant donné que ces variables regroupent plusieurs modalités, telles que par exemple 'Action, Comedy, History' pour 'genres' et 'actor, producer, director' pour 'primaryProfession', nous avons pris la décision de les décomposer en variables binaires, également connues sous le nom de variables dummy.

Cette démarche consiste à créer de nouvelles variables pour chaque modalité possible de 'genres' et 'primaryProfession', où chaque modalité devient une colonne distincte dans notre base de données. Ainsi, si un titre appartient au genre 'Action' par exemple, la variable 'Action' sera représentée par la valeur 1 dans cette colonne, tandis que les autres colonnes auront la valeur 0.

En vue de notre analyse à venir, nous avons pris la décision de restreindre notre ensemble de données aux seules variables suivantes : la durée du film, la public concerné (adulte ou non), la note moyenne donnée par les utilisateurs, le nombre de votants, tous les genres des films ainsi que toutes les professions des réalisateurs présents dans la base.

Pour finir, nous avons structuré les variables sélectionnées en quatre groupes distincts afin de faciliter notre analyse factorielle multiple. Ces groupes sont désignés comme suit : 'Caractéristiques', 'Notes', 'Genres' et 'Professions' (Annexe 4). Chacun de ces regroupements a été constitué pour regrouper et explorer de manière cohérente les aspects spécifiques des données, permettant ainsi une approche plus organisée et précise.

IV. Analyse de données

On observe qu'en moyenne, les films de l'échantillon étudié rassemblent 3525 votants, obtenant une note moyenne d'environ 6. De plus, la durée moyenne des films est d'environ 97 minutes, soit approximativement une heure et trente minutes. (Annexe 5)

Les données de notre base illustrent une image variée et intrigante des films répertoriés. En moyenne, les films reçoivent une note de 6, avec une gamme allant de 1,1 à 10. Cette variation étendue suggère une diversité de goûts et d'appréciations parmi les spectateurs. En explorant plus en détail, il apparaît que 50 % des films obtiennent une note supérieure à 6,2, indiquant une tendance vers des évaluations relativement positives pour une part significative des films.

Par ailleurs, en examinant le nombre de votants, nous constatons une grande disparité : en moyenne, chaque film a été évalué par environ 3525 personnes, avec une diversité marquée dans la participation du public. Certains films ont attiré un nombre considérablement plus élevé de votants, atteignant même jusqu'à 320 606 pour le film 'Les Indestructibles 2'. À l'inverse, un nombre notable des films ont reçu un nombre relativement restreint de votes, allant même aussi bas que 5 pour plus d'une dizaine d'entre eux. De manière intéressante, la moitié des films ont reçu au moins 160 votes, suggérant une participation notable pour une part importante des films, mais mettant également en évidence un écart significatif de participation entre les films les plus populaires et les moins populaires.

En ce qui concerne la durée des films, la moyenne est d'environ 97 minutes, mais cette durée varie considérablement, allant de 11 à 808 minutes. À titre d'exemple, le film 'La Flor' détient le record de la durée la plus longue dans l'histoire du cinéma argentin, avec 808 minutes. Ces variations suggèrent une diversité de formats cinématographiques, des courts métrages aux longs métrages épiques. Notamment, la durée de la moitié des films est d'au moins 94 minutes, démontrant une diversité notable dans les longueurs des films de notre ensemble de données.

Ces constats soulignent la complexité et la diversité des évaluations, de la participation du public et des durées de films au sein de notre base de données cinématographiques, reflétant ainsi la richesse et la variété de l'industrie cinématographique.

Dans notre base de données, nous répertorions un total de 21 genres différents pour les films. Il est important de noter que les films peuvent être associés à plusieurs genres simultanément, ce qui contribue à une diversité de classifications. À la suite de nos analyses, nous remarquons que cinq genres se démarquent davantage que les autres en termes de fréquence. Ces cinq genres dominants sont le "Drame" avec 337 films, suivi de la "Comédie" avec 154 films, des "Documentaires" comptant 121 films, des "Thrillers" avec 64 films, et enfin des films "Romantiques" qui totalisent 60 films (Annexe 6). Ces cinq genres se distinguent nettement des autres par leur fréquence, mettant en évidence leur prépondérance parmi les films répertoriés dans cette base de données.

On observe également les différentes catégories professionnelles auxquelles appartiennent les réalisateurs des films répertoriés dans notre base de données. Par exemple, dans 598 films, le réalisateur est également répertorié comme ayant la catégorie professionnelle de "réalisateur" (director) (Annexe 7). Il est important de noter que ces statistiques reflètent les occurrences où un même individu cumule simultanément plusieurs rôles professionnels pour différents films. Par exemple, Brad Bird exerce à la fois les fonctions de réalisateur, de scénariste, ainsi que d'autres responsabilités relevant de la catégorie miscellaneous. Ainsi, un réalisateur peut être associé à plusieurs catégories professionnelles selon les films, expliquant les divergences entre le nombre total de films et les occurrences pour chaque catégorie professionnelle.

Nous allons réaliser une Analyse Factorielle des Correspondances Multiples (AFM) sur notre ensemble de données qui recense les films de l'année 2018 disponibles sur la plateforme IMDb. Cette analyse sera axée particulièrement sur l'exploration des relations entre les différentes caractéristiques des films tels que les genres, les réalisateurs et d'autres informations pertinentes, mettant spécialement en avant l'importance et l'impact des évaluations attribuées aux films. Notre objectif est de déceler des tendances, des regroupements ou des associations

entre ces variables, en accordant une attention particulière à l'influence des évaluations sur la perception globale des films et leur positionnement dans cette analyse. Ainsi, bien que cette analyse puisse indiquer les tendances générales montrant quels aspects peuvent être associés à des évaluations plus élevées, elle ne détermine pas directement les raisons précises derrière les notes attribuées à un film donné.

À la suite de notre analyse, on observe que les deux premières dimensions extraites expliquent environ 9,2% de la variabilité totale de nos données (Annexe 8). Ces deux aspects principaux que nous explorons capturent une partie de la diversité des caractéristiques des films étudiés, bien que cela ne représente qu'une petite fraction de l'ensemble des informations contenues dans notre base de données. Cela suggère qu'il existe encore une grande variété de facteurs à prendre en compte pour comprendre pleinement ce qui influence les évaluations et les caractéristiques des films sur cette plateforme spécifique.

En observant les données, la première dimension semble moins influencée par les caractéristiques des films ainsi que par les professions des réalisateurs. Cela se reflète par des contributions relativement faibles de ces variables dans la première dimension, suggérant qu'elles ont moins d'impact ou sont moins liées à cette dimension. D'autre part, la deuxième dimension montre des relations plus marquées avec les caractéristiques des films, suggérant une plus grande importance de ces aspects dans la construction de cette dimension. Cependant, elle semble moins influencée par les genres des films par rapport à la première dimension. Cela suggère que, bien que les caractéristiques des films aient un lien plus fort avec la deuxième dimension, les genres semblent exercer une influence moins marquée dans cette dimension par rapport à la première (Annexe 9 et Annexe 10).

En examinant de plus près les axes partiels, nous constatons que la première dimension est principalement influencée par le groupe 'Note' à hauteur de 60 % et présente une influence négative similaire provenant du groupe 'Genres'. Quant à la deuxième dimension, elle est principalement influencée positivement par le groupe 'Genres' à hauteur de 75 %, suivis par le groupe 'Note' avec une influence positive d'environ 50 %. Par ailleurs, il est à noter que le groupe 'Profession' présente une influence relativement faible, que ce soit sur la première ou la deuxième dimension (Annexe 11).

Il est intéressant de noter que les films documentaires ont tendance à recevoir des évaluations remarquablement positives de la part du public. Leur capacité à présenter des faits, à informer et à éduquer semble susciter un fort engouement et à générer des critiques favorables, ce qui se reflète dans des notes globalement élevées attribuées à ce genre de films. Par contraste, les films d'horreur ont tendance à enregistrer des évaluations basses. La nature souvent intense, anxiogène voire choquante de ce genre peut provoquer des réactions mitigées voire des critiques plus sévères de la part des spectateurs, se traduisant ainsi par des notes relativement basses. De plus, une tendance intéressante émerge quant à la participation du public : les films d'animations, d'aventure, d'action et de fantaisie semblent attirer un nombre plus élevé de votants. Ce constat peut s'expliquer par l'attrait généralement large et l'engagement émotionnel que ces genres offrent, suscitant ainsi un intérêt plus marqué et une participation accrue des spectateurs pour exprimer leur avis et leur appréciation (Annexe 12).

En outre, nous constatons que les individus sont répartis de manière assez homogène autour de la moyenne. Ainsi, cela suggère que les films présentent une certaine similarité à travers différentes caractéristiques (Annexe 13).

V. Conclusion

En conclusion, ces tendances observées à travers l'analyse des données soulignent les préférences marquées du public pour certains genres cinématographiques et mettent en lumière l'impact significatif du contenu et du style des films sur leur réception et l'engagement du public.

L'exploration approfondie des films de l'année 2018 sur la plateforme IMDb a révélé des tendances intéressantes dans les préférences du public et les caractéristiques des films. Nous avons noté que les films documentaires bénéficient généralement d'évaluations positives, probablement en raison de leur capacité à informer et à éduquer. En revanche, les films d'horreur obtiennent généralement des notes plus basses en raison de leur nature intense et parfois choquante pour les spectateurs.

De plus, les films d'animations, d'aventure, d'action et de fantaisie attirent un nombre plus conséquent de votants, ce qui suggère un attrait plus large et un engagement émotionnel fort pour ces genres.

Nous avons également observé une répartition homogène des films autour de la moyenne, indiquant une certaine similarité entre eux malgré leurs diverses caractéristiques.

En résumé, cette analyse des données IMDb pour l'année 2018 offre un aperçu des préférences du public et des particularités des films, mettant en évidence la diversité des évaluations, des genres et de la participation du public dans l'industrie cinématographique durant cette période spécifique.

PARTIE B : Annexes

➢ Annexe 1 : Lien vers l'open data IMDb

Ce lien https://developer.imdb.com/non-commercial-datasets/ vous dirige vers la plateforme IMDb qui contient une multitude de bases de données fournissant des informations variées, telles que les évaluations, les noms des réalisateurs, les acteurs, les équipes de production, les résumés des films, et bien d'autres détails relatifs aux films, émissions de télévision et séries répertoriés sur la plateforme.

➢ Annexe 2 : Modèle Conceptuel de Données

Nous avons fusionné les bases de données "title.basics" et "title.ratings" en utilisant la clé "tconst", puis fusionné ces données avec "title.principals" en utilisant la même clé, enfin nous avons fusionné ces données avec "name.basics" en utilisant la clé "nconst" que nous renommerons plus tard par "directors" .

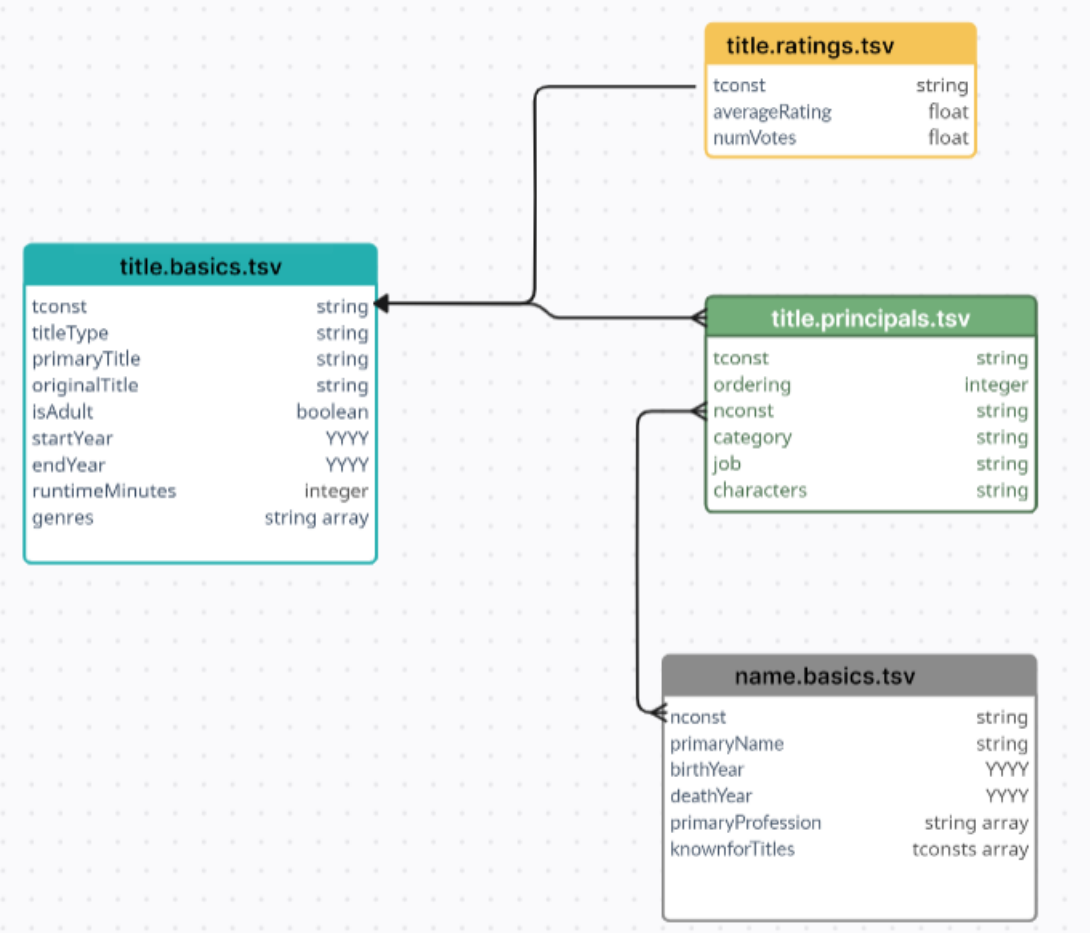


Figure 3 : Modèle Conceptuel de Données

➢ Annexe 3 : Dictionnaire des variables

La base de données finale est composée de 1292 titres définis par 13 variables :

Variables	Description	Type de variable
tconst	Identifiant alphanumérique unique du titre	Chaîne de caractère
title	Le nom du titre	Chaîne de caractère
genres	Le(s) genre(s) du titre (par exemple Action, Comédie)	Chaîne de caractère
titleType	Le type du titre (par exemple film, série)	Chaîne de caractère
runtimeMinutes	Durée d'exécution du titre, en minutes	Numérique
isAdult	Film pour les + 18 ans ou non (0 : non adulte ; 1 : adulte)	Binaire
averageRating	Moyenne pondérée de toutes les notes individuelles données par les utilisateurs	Numérique
numVotes	Nombre de votes reçus par le titre	Numérique
directors	Identifiant alphanumérique unique du réalisateur principal du titre	Chaîne de caractère
primaryName	Le nom du réalisateur	Chaîne de caractère
birthYear	L'année de naissance du réalisateur	Date
deathYear	L'année de décès du réalisateur (NA s'il n'est pas décédé)	Date
primaryProfession	Les 3 principales professions du réalisateur (par exemple acteur, producteur)	Chaîne de caractère

(Le code couleur étant la provenance des variables selon les bases de données : voir Annexe 2)

➢ Annexe 4 : Répartition des groupes

Cette structuration a été réalisée dans le but de regrouper les variables selon des catégories spécifiques pour faciliter l'analyse factorielle multiple :

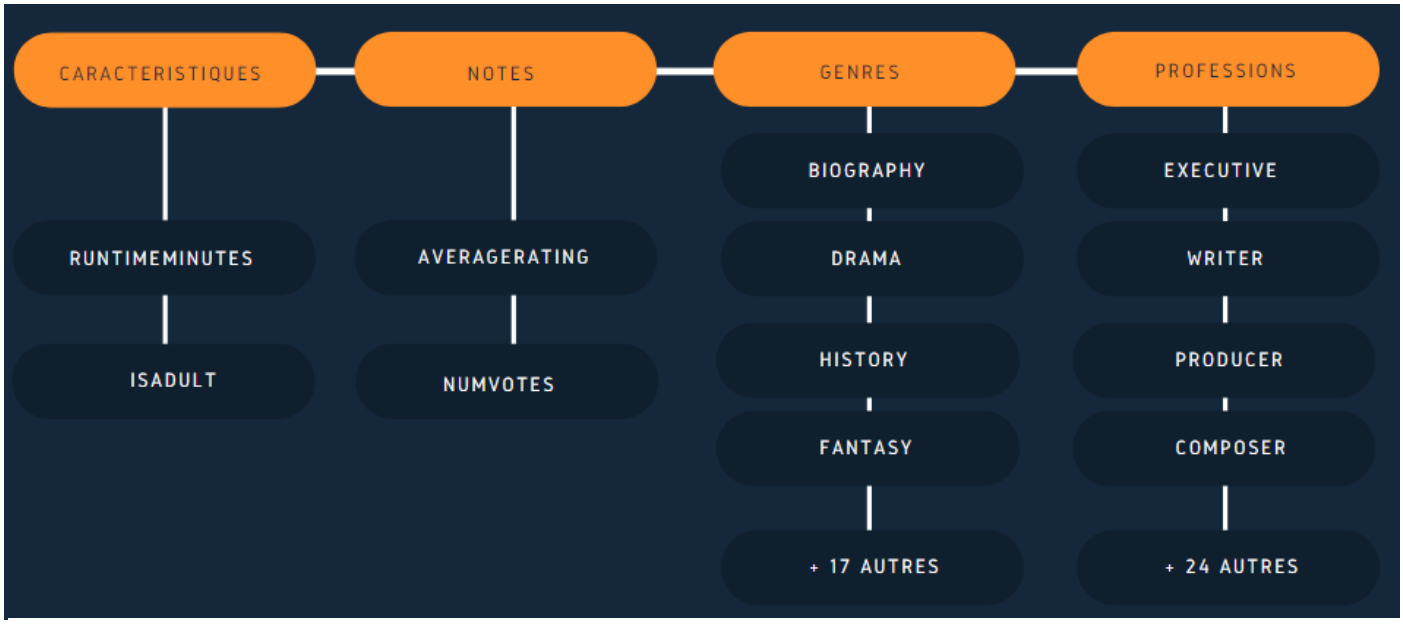


Figure 4 : Mise en groupe des variables présentent dans la base de données

➢ Annexe 5 : Résumé statistique

Un résumé statistique des différentes variables contenues dans notre base de données, telles que la note des films, le nombre de votants, ainsi que la durée des films.

```
> summary(newDF[c("averageRating", "numVotes", "runtimeMinutes")])
 averageRating      numVotes         runtimeMinutes
 Min.   : 1.100   Min.   :     5.0   Min.   : 11.00
 1st Qu.: 5.400   1st Qu.:    35.5   1st Qu.: 83.00
 Median : 6.200   Median :   160.0   Median : 94.00
 Mean   : 6.077   Mean   :  3525.1   Mean   : 97.72
 3rd Qu.: 7.000   3rd Qu.:   744.0   3rd Qu.:105.00
 Max.   :10.000   Max.   :320606.0   Max.   :808.00
```

➢ Annexe 6 : Genres les plus fréquents

Le graphe ci-dessous représente les cinq genres les plus fréquents dans notre base de données. La sélection de ces cinq genres s'explique par la présence d'un nombre conséquent de genres disponibles (21).

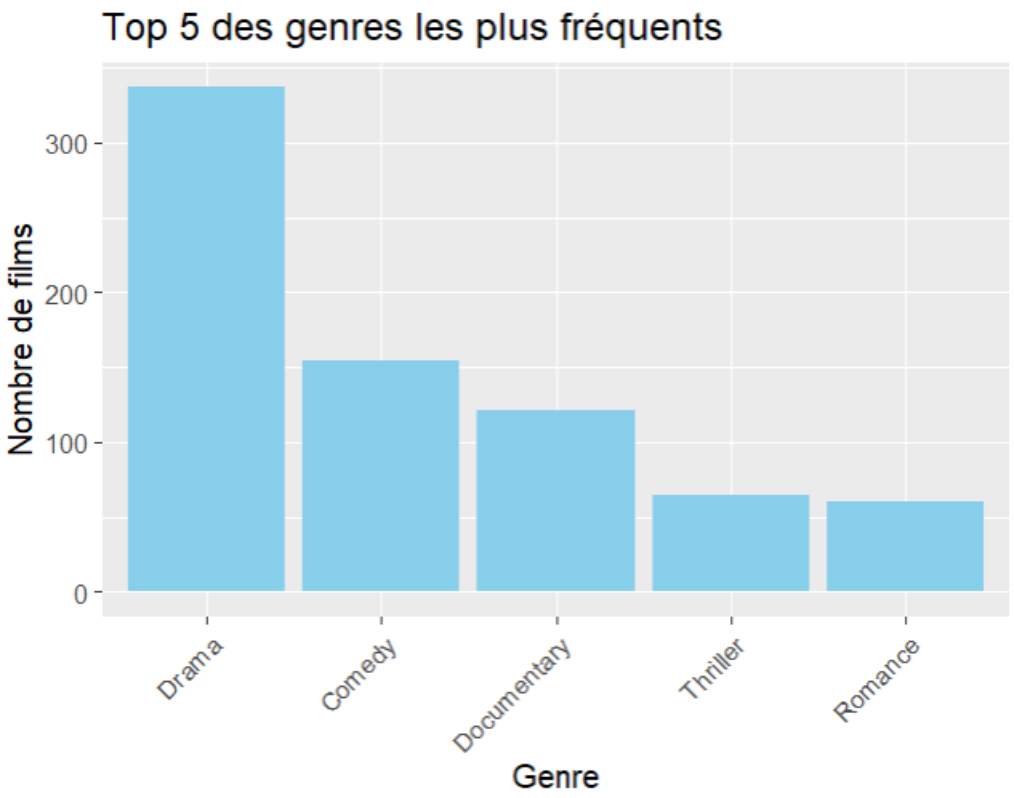


Figure 5 : Les 5 genres les plus fréquents dans la base de données
Lecture : Le genre « Comédie » est associé pour environ 150 films.

➢ Annexe 7 : Catégorie professionnelles les plus fréquents

Le graphe ci-dessous représente les cinq catégories les plus fréquents correspondant aux réalisateurs dans notre base de données. La sélection de ces cinq catégories s'explique par la présence d'un nombre conséquent de catégories disponibles (28).

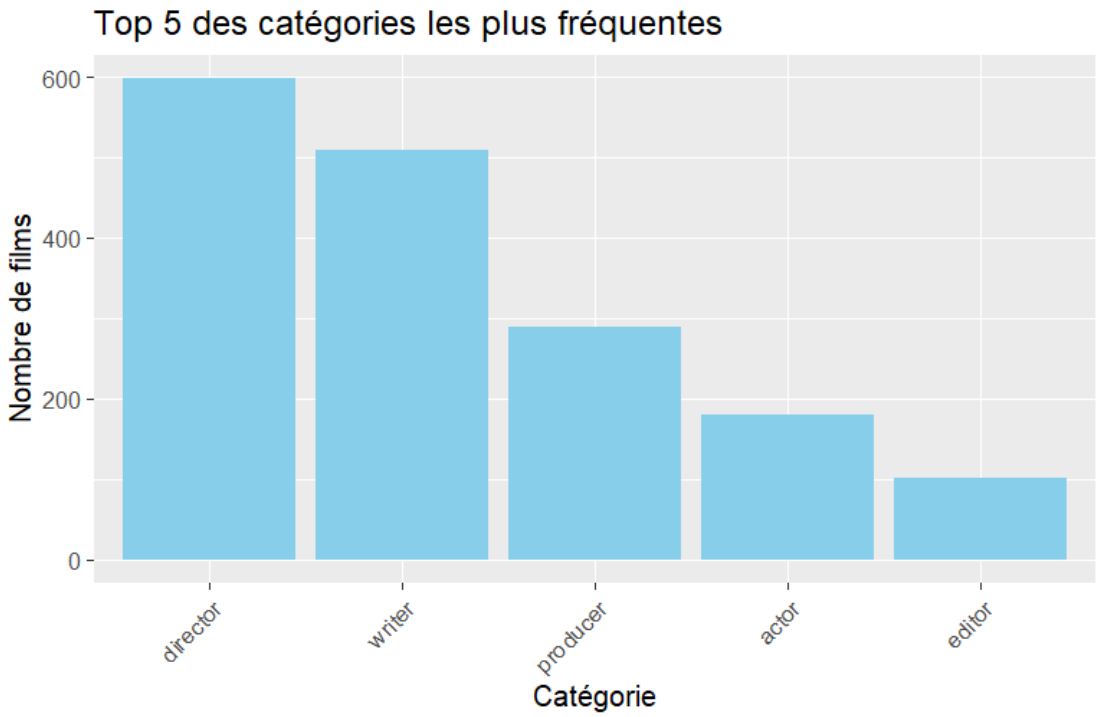


Figure 6 : Les 5 catégories les plus fréquentes dans la base de données

Lecture : 600 des films ont un réalisateur qui a comme catégorie « Réalisateur », ce qui correspond à environ la moitié des films dans notre base de données.

➢ Annexe 8 : Variance expliquée par chaque dimension dans l'AFM

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9
Variance	1.455	1.332	1.203	1.075	1.026	0.930	0.913	0.898	0.864
% of var.	4.813	4.407	3.981	3.557	3.393	3.075	3.019	2.969	2.859
Cumulative % of var.	4.813	9.219	13.200	16.758	20.151	23.226	26.245	29.214	32.073
	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14	Dim.15	Dim.16	Dim.17	Dim.18
Variance	0.818	0.752	0.741	0.723	0.713	0.669	0.662	0.653	0.637
% of var.	2.706	2.488	2.450	2.393	2.358	2.215	2.191	2.161	2.106
Cumulative % of var.	34.779	37.267	39.718	42.111	44.469	46.684	48.875	51.035	53.142
	Dim.19	Dim.20	Dim.21	Dim.22	Dim.23	Dim.24	Dim.25	Dim.26	Dim.27
Variance	0.626	0.607	0.595	0.582	0.572	0.565	0.562	0.551	0.546
% of var.	2.070	2.007	1.969	1.925	1.893	1.870	1.860	1.824	1.808
Cumulative % of var.	55.211	57.218	59.186	61.111	63.005	64.875	66.735	68.559	70.366
	Dim.28	Dim.29	Dim.30	Dim.31	Dim.32	Dim.33	Dim.34	Dim.35	Dim.36
Variance	0.538	0.531	0.525	0.508	0.491	0.462	0.456	0.443	0.435
% of var.	1.780	1.758	1.736	1.682	1.624	1.528	1.508	1.464	1.440
Cumulative % of var.	72.146	73.904	75.640	77.322	78.946	80.474	81.982	83.446	84.885
	Dim.37	Dim.38	Dim.39	Dim.40	Dim.41	Dim.42	Dim.43	Dim.44	Dim.45
Variance	0.412	0.404	0.396	0.386	0.373	0.362	0.355	0.320	0.308
% of var.	1.364	1.336	1.310	1.276	1.234	1.198	1.174	1.059	1.020
Cumulative % of var.	86.250	87.586	88.896	90.172	91.405	92.604	93.777	94.836	95.857
	Dim.46	Dim.47	Dim.48	Dim.49	Dim.50	Dim.51	Dim.52	Dim.53	
Variance	0.279	0.270	0.219	0.190	0.153	0.119	0.022	0.000	
% of var.	0.922	0.894	0.726	0.628	0.508	0.392	0.073	0.000	
Cumulative % of var.	96.779	97.673	98.399	99.027	99.534	99.927	100.000	100.000	

Figure 7 : Variance expliquée par chaque dimension dans l'AFM

Lecture : Les dimensions 1 et 2 expliquent à elles seules environ 9,2% de la variance.

Dans notre base de données nous disposons de 50 variables binaires et 3 variables non binaires, la nécessité de 53 dimensions pour expliquer 100 % de la variance peut être expliquée par la redondance d'information.

Même si nous avons 50 variables binaires, elles peuvent ne pas être totalement indépendantes. Si certaines de ces variables sont corrélées entre elles, cela peut nécessiter davantage de dimensions pour capturer l'ensemble de l'information. Par exemple, si certaines variables binaires sont fortement corrélées elles n'apportent pas d'information unique et peuvent être considérées comme redondantes.

➢ Annexe 9 : Contribution des groupes aux dimensions dans l'AFM

L'annexe suivante détaille les contributions des différents groupes analysés dans les dimensions extraites via l'AFM. Les valeurs de contribution (CTR) et de qualité de la représentation (COS2) sont présentées pour chaque variable dans les deux premières dimensions (Dim.1, Dim.2), fournissant ainsi des indications sur l'importance et la représentation de ces aspects dans chaque dimension de l'AFM.

Groups

	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Caracteristiques	0.071	4.847	0.005	0.216	16.239	0.047
Note	0.637	43.775	0.239	0.300	22.554	0.053
Genres	0.555	38.135	0.039	0.631	47.368	0.050
Profession	0.193	13.243	0.004	0.184	13.839	0.004

Figure 8 : Contribution des groupes aux dimensions dans l'AFM

Lecture : Le groupe des caractéristiques a une contribution de 4,8% dans la dimension 1 avec une qualité de représentation de 0.005.

➢ Annexe 10 : Visualisation de la contribution des groupes

Dans notre analyse, nous avons choisi de conserver tous les groupes comme actifs.

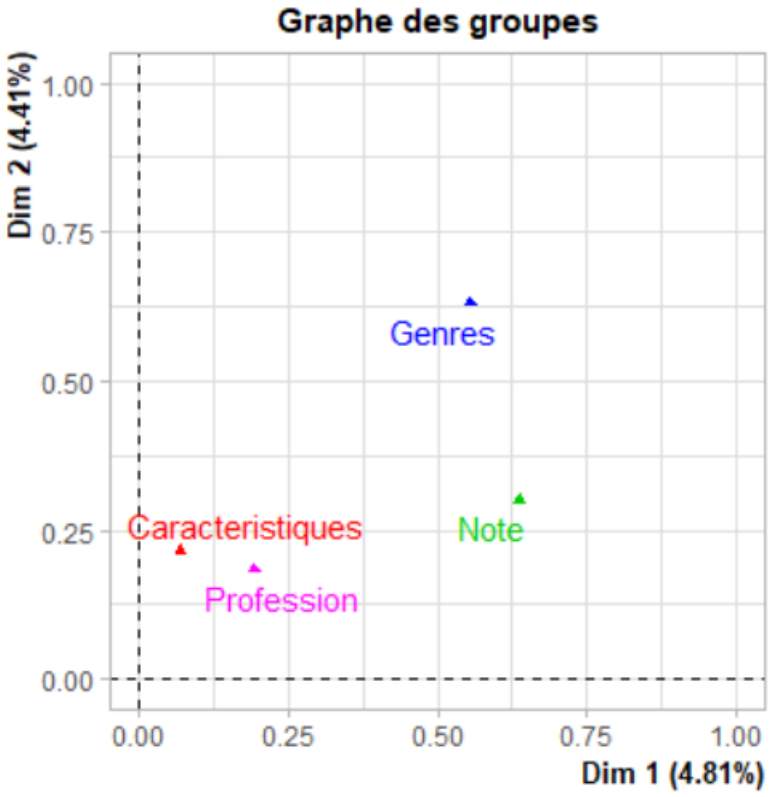


Figure 9 : Graphique des groupes de l'AFM

Lecture : Le groupe le plus représentatif de la dimension 1 est le groupe Note regroupant les variables des évaluations et du nombre de votants.

➢ Annexe 11 : Visualisation des axes partiels

L'annexe suivante présente les graphiques des axes partiels issus de l'AFM. Ces graphiques représentent visuellement les contributions des différentes variables à la construction des dimensions principales dans l'AFM. Les axes partiels offrent une perspective sur l'importance relative des variables ou des catégories de variables dans la formation de chaque dimension de l'analyse.

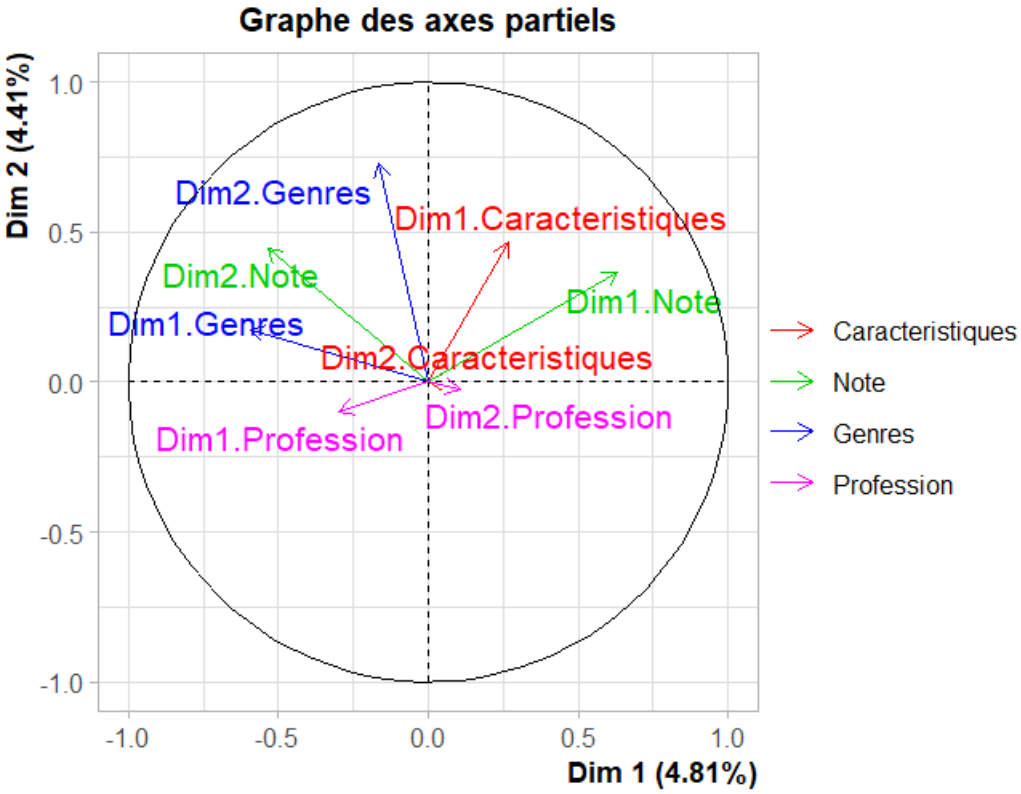
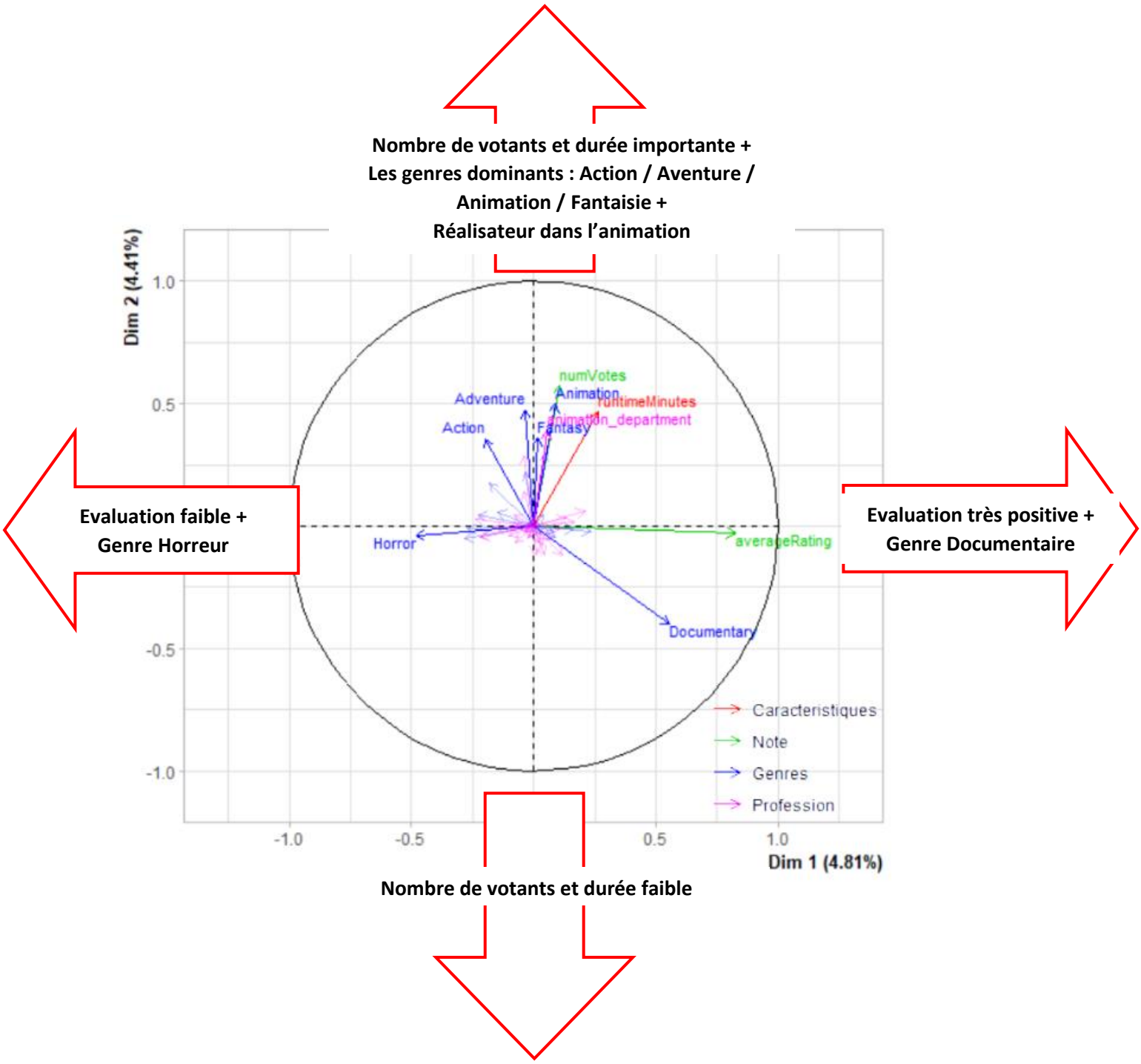


Figure 10 : Graphique des axes partiels de l'AFM

Lecture : Le groupe Genres est celui qui contribue le plus à la construction de la deuxième dimension, à hauteur de 75%.

➢ Annexe 12 : Cercle des corrélations de l'AFM



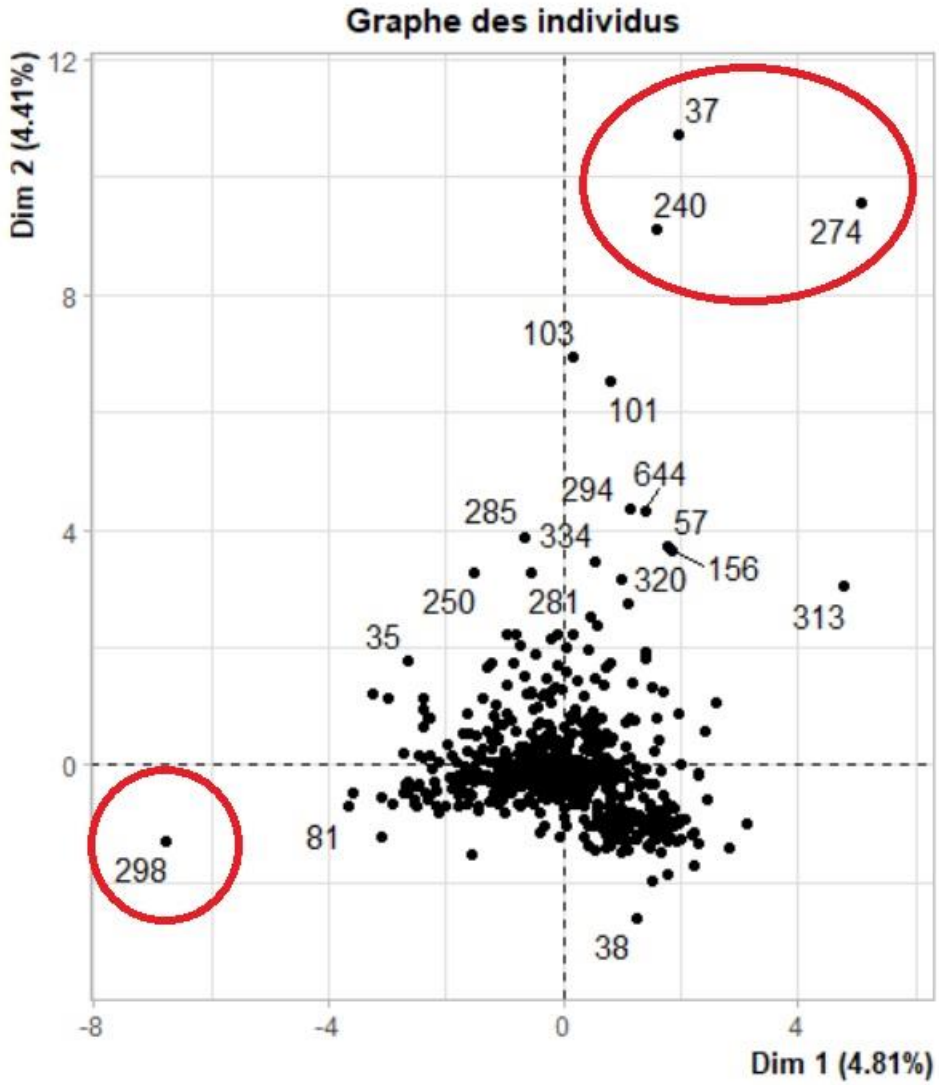
➢ Annexe 13 : Graphique des individus de l'AFM

Sur ce graphique, nous remarquons la présence d'individus atypiques, ces derniers se démarquent pour différentes raisons :

Les individus 37 et 240 : Ils se distinguent par un nombre de votants considérablement plus élevé que les autres films (+ de 300 000).

L'individu 274 : Il se distingue par une durée bien plus longue que les autres films, avec une durée de 808 minutes.

L'individu 298 : Il se démarque en étant le seul film de notre base ayant un réalisateur qui a pour caractéristique les effets spéciaux.



PARTIE C : Codes

➔ R Studio (Préparation de la base finale)

```
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importer les librairires nécessaires -----
library(disk.frame)
library(dplyr)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importation et manipulations de title_basics -----
# Importer le fichier tsv title_basics dans un disk.frame
title_basics <- as.disk.frame(read.delim("data/title_basics.tsv", sep = "\t"))
# Remplacer "\N" par NA dans toutes les colonnes
title_basics <- title_basics %>%
  mutate_all(~na_if(., "\\N"))
# Supprimer les lignes avec des valeurs manquantes dans title_basics
title_basics_clean <- na.omit(title_basics)
# Filtrer la base de données en fonction de la variable startYear égale à 2018
title_basics_clean <- filter(title_basics_clean, startYear == 2018)
# Convertir le disk.frame en data.frame
data_filtered_basics_df <- as.data.frame(title_basics_clean)
# Exporter le data.frame en CSV
write.csv(data_filtered_basics_df, "data/data_filtered_basics.csv", row.names = FALSE)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importation et manipulations de title_ratings -----
# Importer le fichier tsv title_ratings dans un disk.frame
title_ratings <- as.disk.frame(read.delim("data/title_ratings.tsv", sep = "\t"))
# Remplacer "\N" par NA dans toutes les colonnes
title_ratings <- title_ratings %>%
  mutate_all(~na_if(., "\\N"))
# Supprimer les lignes avec des valeurs manquantes dans title_ratings
title_ratings_clean <- na.omit(title_ratings)
# Convertir le disk.frame en data.frame
data_filtered_ratings_df <- as.data.frame(title_ratings_clean)
# Exporter le data.frame en CSv
write.csv(data_filtered_ratings_df, "data/data_filtered_ratings.csv", row.names = FALSE)
# Fusion des deux dataframes basics et ratings -----
# Fusion des dataframes basics et ratings en utilisant la variable 'tconst'
merged_data <- merge(data_filtered_ratings_df, data_filtered_basics_df, by = 'tconst')
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importation et manipulations de title_akas -----
# Importer le fichier tsv title_akas dans un disk.frame
title_akas <- as.disk.frame(read.delim("data/title_akas.tsv", sep = "\t"))
# Remplacer "\N" par NA dans toutes les colonnes
title_akas <- title_akas %>%
  mutate_all(~na_if(., "\\N"))
# Supprimer les lignes avec des valeurs manquantes dans title_akas
title_akas_clean <- na.omit(title_akas)
# Filtrer la base de données en fonction de la variable isOriginalTitle égale à 1
title_akas_clean <- filter(title_akas_clean, isOriginalTitle == 1)
```

```
# Convertir le disk.frame en data.frame
data_filtered_akas_df <- as.data.frame(title_akas_clean)
# Exporter le data.frame en CSv
write.csv(data_filtered_akas_df, "data/data_filtered_akas.csv", row.names = FALSE)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Fusion des deux dataframes merged_data et akas -----
# Fusion des dataframes merged_data et akas en utilisant la variable 'tconst' pour merged_data et titleId pour akas
merged_data2 <- merge(merged_data, data_filtered_akas_df, by.x = 'tconst', by.y = 'titleId')
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importation et manipulations de title_crew -----
# Importer le fichier tsv title_crew dans un disk.frame
title_crew <- as.disk.frame(read.delim("data/title_crew.tsv", sep = "\t"))
# Remplacer "\N" par NA dans toutes les colonnes
title_crew <- title_crew %>%
  mutate_all(~na_if(., "\\N"))
# Supprimer les lignes avec des valeurs manquantes dans title_crew
title_crew_clean <- na.omit(title_crew)
# Convertir le disk.frame en data.frame
data_filtered_crew_df <- as.data.frame(title_crew_clean)
# Exporter le data.frame en CSv
write.csv(data_filtered_crew_df, "data/data_filtered_crew.csv", row.names = FALSE)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Fusion des deux dataframes merged_data2 et crew -----
# Fusion des dataframes merged_data2 et crew en utilisant la variable 'tconst'
merged_data3 <- merge(merged_data2, data_filtered_crew_df, by = 'tconst')
# Filtrer les lignes en ne gardant que les lignes ayant qu'un seul directors et writers, et enlever celles n'en ayant pas.
merged_data3 <- merged_data3 %>%
  filter(!grepl(",", directors) & !grepl(",", writers) & !is.na(directors) & !is.na(writers) & writers != "\\N" & directors !=
"\\N")
# Enlever la variable ordering et writers
merged_data3 <- merged_data3 %>%
  select(-ordering, -writers)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importation et manipulations de title_principals -----
# Importer le fichier tsv title_principals dans un disk.frame
title_principals <- as.disk.frame(read.delim("data/title_principals.tsv", sep = "\t"))
# Remplacer "\N" par NA dans toutes les colonnes
title_principals <- title_principals %>%
  mutate_all(~na_if(., "\\N"))
# Supprimer les lignes avec des valeurs manquantes dans title_principals
title_principals_clean <- na.omit(title_principals)
# Convertir le disk.frame en data.frame
data_filtered_principals_df <- as.data.frame(title_principals_clean)
# Exporter le data.frame en CSv
write.csv(data_filtered_principals_df, "data/data_filtered_principals.csv", row.names = FALSE)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importation et manipulations de name_basics -----
# Importer le fichier tsv name_basics dans un disk.frame
name_basics <- as.disk.frame(read.delim("data/name_basics.tsv", sep = "\t"))
# Remplacer "\N" par NA dans toutes les colonnes
name_basics <- name_basics %>%
  mutate_all(~na_if(., "\\N"))
```

```
# Supprimer les lignes avec des valeurs manquantes dans name_basics
name_basics_clean <- na.omit(name_basics)
# Convertir le disk.frame en data.frame
data_filtered_nbasics_df <- as.data.frame(name_basics_clean)
# Exporter le data.frame en CSv
write.csv(data_filtered_nbasics_df, "data/data_filtered_nbasics.csv", row.names = FALSE)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Fusion des deux dataframes merged_data3 et name_basics -----
# Fusion des dataframes merged_data3 et name_basics en utilisant la variable 'tconst'
merged_data4 <- merge(merged_data3, data_filtered_nbasics_df, by.x = "directors", by.y = "nconst", all.x = TRUE)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Mettre en évidence les valeurs manquantes ------
# Remplacer \N par NA dans votre dataframe
final_merged_data <- data.frame(lapply(merged_data4, function(x) {
  x[x == "\\N"] <- NA
  return(x)
}))
# Suppressions des variables non utiles  ------
final_merged_data <- subset(final_merged_data, select = -c(primaryTitle, originalTitle, startYear, endYear, region,
language, types, attributes, isOriginalTitle, knownForTitles))
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Fonction pour extraire le premier bloc avant la virgule
extract_first_block <- function(value) {
  if (!is.na(value)) {
    # Recherche du premier bloc avant la virgule
    extracted_block <- strsplit(value, ",")[[1]][1]
    return(extracted_block)
  }
  return('')
}
# Appliquer la fonction d'extraction à la colonne 'genre'
final_merged_data$FirstGenre <- sapply(final_merged_data$genres, extract_first_block)
# Organisation des variables ------
final_merged_data <- final_merged_data %>%
  select(tconst, title, FirstGenre, genres, titleType, runtimeMinutes, isAdult, averageRating, numVotes, directors,
primaryName, birthYear, deathYear, primaryProfession)
# Supprimer les lignes contenant des valeurs manquantes (NA) dans toutes les variables sauf "deathYear"
final_merged_data <- final_merged_data[complete.cases(final_merged_data[, !colnames(final_merged_data) %in%
"deathYear"]), ]
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Exporter le data frame final_merged_data en fichier CSV -----
write.csv(final_merged_data, file = "data/data.csv", row.names = FALSE)
```

➔ R Studio (analyse)

```
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importer les librairires nécessaires -----
library(Factoshiny)
library(tidyr)
library(dplyr)
library(ggplot2)
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Importer la base de donnees data.csv -----
data <- read.csv("data/data.csv")
data <- subset(data, select = -c(deathYear))
df <- as.data.frame(data)

df_movies <- subset(df, grepl("^movie$", titleType, ignore.case = TRUE))
df_movies <- subset(df_movies, select = -c(tconst, title, FirstGenre, titleType, directors, primaryName))

# Séparer les genres en différentes colonnes
df_movies_dummies <- df_movies %>%
  separate_rows(genres, sep = ",") %>%  # Séparation des genres
  mutate(genre_present = 1) %>%  # Ajout d'une colonne indiquant la présence du genre
  pivot_wider(names_from = genres, values_from = genre_present, values_fill = 0)  # Pivoter pour obtenir les
dummies

df_movies_dummies[is.na(df_movies_dummies)] <- 0

# Séparer les professions en différentes colonnes
df_movies_dummies <- df_movies_dummies %>%
  separate_rows(primaryProfession, sep = ",") %>%  # Séparation des professions
  mutate(profession_present = 1) %>%  # Ajout d'une colonne indiquant la présence de la profession
  pivot_wider(names_from = primaryProfession, values_from = profession_present, values_fill = 0)  # Pivoter pour
obtenir les dummies

df_movies_dummies[is.na(df_movies_dummies)] <- 0

cor(df_movies_dummies)
#On observe que les variables sont très faiblement corrélées dans la globalité, donc il y a peu de chance que les
variables aillent dans le même sens.

#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-

# Analyse descriptive ---
newDF <-
df_movies_dummies[,c("runtimeMinutes","isAdult","averageRating","numVotes","Biography","Drama","History","F
antasy","Music","Comedy","Documentary","Horror","Mystery","Sport","Crime","Thriller","Action","Adventure","Ro
mance","Animation","Family","Sci-
Fi","Musical","Western","War","special_effects","make_up_department","script_department","editorial_departme
nt","casting_department","executive","sound_department","camera_department","music_department","productio
n_manager","assistant_director","writer","producer","music_artist","art_director","production_designer","cinemat
ographer","animation_department","visual_effects","stunts","composer","editor","soundtrack","actor","director","
art_department","actress","miscellaneous")]
```

```
summary(newDF[c("averageRating", "numVotes", "runtimeMinutes")])

# Comptage du nombre de lignes pour chaque genre
genre_counts <- colSums(newDF[, c("Biography", "Drama", "History", "Fantasy", "Music", "Comedy",
"Documentary", "Horror", "Mystery", "Sport", "Crime", "Thriller", "Action", "Adventure", "Romance", "Animation",
"Family", "Sci-Fi", "Musical", "Western", "War")])
# Affichage du nombre de lignes pour chaque genre
print(genre_counts)
# Création d'un dataframe pour le graphique
genre_data <- data.frame(Genre = names(genre_counts), Count = genre_counts)
# Sélection des cinq genres les plus représentés
top5_genres <- head(genre_data[order(-genre_data$Count), ], 5)
# Création du graphique pour le top 5
ggplot(top5_genres, aes(x = reorder(Genre, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Top 5 des genres les plus fréquents", x = "Genre", y = "Nombre de films") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Comptage du nombre de lignes pour chaque catégorie
category_counts <- colSums(newDF[,
c("special_effects","make_up_department","script_department","editorial_department","casting_department","ex
ecutive","sound_department","camera_department","music_department","production_manager","assistant_direct
or","writer","producer","music_artist","art_director","production_designer","cinematographer","animation_depart
ment","visual_effects","stunts","composer","editor","soundtrack","actor","director","art_department","actress","mi
scellaneous")])
# Affichage du nombre de lignes pour chaque category
print(category_counts)
# Création d'un dataframe pour le graphique
category_data <- data.frame(Category = names(category_counts), Count = category_counts)
# Sélection des cinq genres les plus représentés
top5_category <- head(category_data[order(-category_data$Count), ], 5)
# Création du graphique pour le top 5
ggplot(top5_category, aes(x = reorder(Category, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Top 5 des catégories les plus fréquentes", x = "Catégorie", y = "Nombre de films") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##--##--###--##-
# Analyse Factorielle Multiple (AFM) ----
res.MFA<-MFA(newDF,group=c(2,2,21,28),
type=c("s","s","s","s"),name.group=c("Caracteristiques","Note","Genres","Profession"),graph=FALSE)
summary(res.MFA)

# Graphe des groupes
plot.MFA(res.MFA, choix="group",title="Graphe des groupes")
# Graphe des axes partiels
plot.MFA(res.MFA, choix="axes",title="Graphe des axes partiels",habillage='group', shadow=TRUE)
# Cercle des corrélations
plot.MFA(res.MFA, choix="var",habillage='group',title="Cercle des corrélations", cex=0.7, select="contrib 10",
shadow=TRUE)
# Graphe des individus
plot.MFA(res.MFA, choix="ind",lab.par=FALSE,title="Graphe des individus")
```