ANALYSE DE L'ONGLET TENDANCE DE YOUTUBE

MAT5052

Analyse exploratoire

Submitted To :
Myriam Bertrand
Emmanuelle Clayes

Submitted By :
Léonard Benedetti
Joris Garnier
Maxime Sazadaly

Résumé

Le présent rapport est une étude de cas ayant trait à l'analyse exploratoire d'un jeu de données concernant les vidéos se retrouvant dans l'onglet Tendance de Youtube. Pour étudier les liaisons entre les différentes variables à notre disposition caractérisant ces vidéos, nous avons tenté décidé d'utiliser une Analyse en Composantes Principales, une Analyse Factorielle Multiple et une Classification Hiérarchique Ascendante.

1 Présentation du dataset - Youtube Trending Videos

Le dataset utilisé pour notre projet est le dataset *Youtube Trending Videos* provenant de la base de Kaggle. Il réunit des statistiques journalières relatives aux vidéos présentes dans l'onglet *Tendance* de Youtube dans deux pays, à savoir les États-Unis et la Grande-Bretagne.

Ces vidéos sont ainsi décrites par 9 variables où $title$ est le nom de la vidéo, $channel\_title$ est le titre de la chaîne ayant produit la vidéo, $category\_id$ est la catégorie dans laquele la vidéo a été classée, $tags$ est une concaténation de tag associés à la vidéo, $views$ est le nombre de vues de la vidéo, $likes$ est le nombre de likes de la vidéo, $dislikes$ est le nombre de dislikes de la vidéo, $comment\_total$ est le nombre de commentaires liés à la vidéo et $date$ est la date à laquelle la vidéo s'est retrouvé dans l'onglet *Tendance*. D'autres informations telles que l'identifiant de la vidéo ou sa vignette étaient mises à disposition mais n'ont pas été utilisées.

De même, un second jeu de données regroupant l'ensemble des commentaires de ces vidéos étaient mis à disposition. Son utilisation sera discutée plus tard dans l'étude.

```
> head(videos_GB.quanti)
     views likes dislikes comment_total
1  7426393 78240    13548           705
2   494203  2651     1309             0
3   142819 13119      151          1141
4  1580028 65729     1529          3598
5    40592  5019       57           490
6   317696  9449      135           464
```

2 Augmentation du dataset

Le dataset choisi était composé de près de 13997 vidéos, 6997 pour la Grande-Bretagne et 7000 pour les États-Unis. Pour rendre l'étude de celui-ci plus intéressant, nous avons décidé de rajouter des variables permettant de décrire certaines variables qualitatives de façon quantitative.

Nous avons ainsi décidé d'introduire, pour chaque vidéo, les variables suivantes : $lenTitle$, le nombre de caractères composant le titre ; $nbCap$, le nombre de majuscules composant le titre ; $nbLow$, le nombre de minuscules composant le titre ; $nb\_days\_trending$, le nombre de jours où la vidéo a été présente dans l'onglet Tendance de YouTube.

Il aurait été également possible d'introduire d'autres variables caractérisant les commentaires (longueur moyenne des commentaires, fréquence d'émojis, polarité du commentaire, etc.) et d'ajouter cela à l'analyse. Néanmoins nous avons préféré ne pas le faire pour éviter d'introduire un biais dans les données. A contrario, nous estimons que les variables que nous avons introduites n'en ajoute pas, celles-ci étant directement extraites d'autres variables déjà présentes.

3 Répartition des données

Les vidéos étudiées sont décrites par 12 variables une fois notre augmentation effectuée. Parmi celles-ci, on trouve 4 variables qualitatives et 8 variables quantitatives. Avant d'explorer les données de façon plus poussée, nous avons décidé d'analyser celles-ci de façon plus générale.

3.1 Word cloud

Nous avons tout d'abord voulu repérer quels étaient les thèmes récurrent des vidéos étudiées. Pour cela, la génération d'un nuage de mot relatif aux titres des vidéos nous a semblé particulièrement adapté. Pour pouvoir obtenir des résultats significatifs, nous avons d'abord élagué ces titres en retirant la ponctuation et les mots fréquents de la langue anglaise (the, I, should, etc.)

(a) GB

(b) US

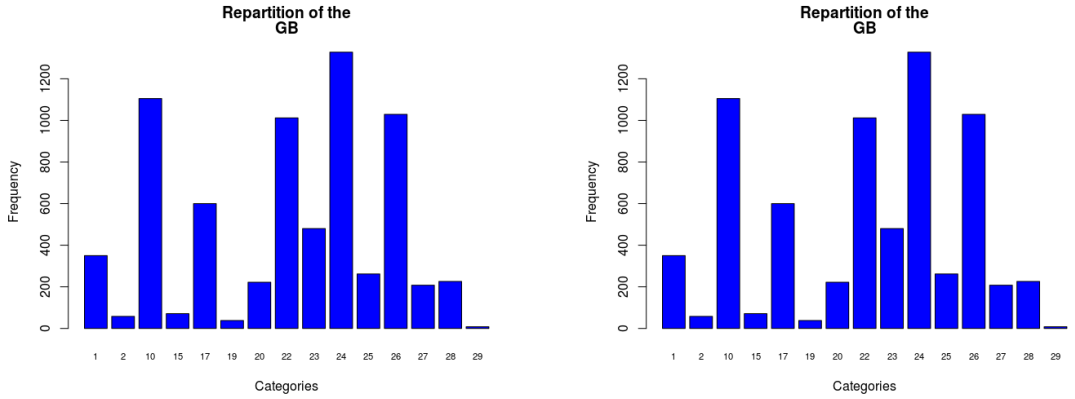
FIGURE 1 – Nuage de mots composant le titre des vidéos

```
generate_wordcloud <- function(variable, max_words = 100) {

  titlesCorpus <- Corpus(VectorSource(variable))
  titlesCorpus <- tm_map(titlesCorpus, PlainTextDocument)
  titlesCorpus <- tm_map(titlesCorpus, tolower)
  titlesCorpus <- tm_map(titlesCorpus, removePunctuation)
  titlesCorpus <- tm_map(titlesCorpus, removeWords,
                         stopwords('english'))
  titlesCorpus <- iconv(titlesCorpus, to = "utf-8")

  wordcloud(titlesCorpus, max.words = max_words,
            colors = brewer.pal(10, "Spectral"))

}
```

On peut ainsi voir que, dans les deux cas, les mots les plus présents sont "Official" et "video".

3.2 Barplots

Nous avons ensuite décidé de nous intéresser à la répartition des vidéos entre les différentes catégories. Pour cela, nous avons décidé de générer les barplots à partir de la variable $category\_id$, que voici :

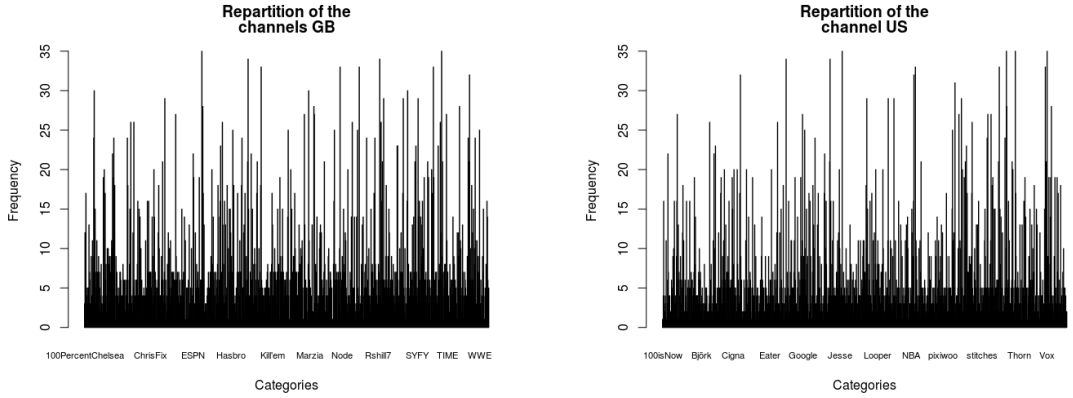


(a) GB (b) US

FIGURE 2 – Barplot - Répartition des catégories parmi les vidéos

La catégorie la plus représentée est ainsi **Entertainment** aussi bien en Grande-Bretagne qu'aux USA, suivie par **Music**. Bien que la répartition des videos soient similaires entre les deux pays étudiés, on peut constater que les vidéos traitant de **People & Blogs** et **News et Politics** se retrouvent significativement plus souvent dans l'onglet Tendance en Grande-Bretagne qu'aux États-Unis.

Nous avons également tenté de générer un barplot relatif aux chaînes produisant les vidéos. Néanmoins, étant donné le nombre de modalités possibles pour la variable $channel\_title$, le résultat est illisible. Il est cependant notable que certaines chaînes sont significativement plus représentées.



(a) GB (b) US

FIGURE 3 – Nombre de vidéos dans l'onglet de chaque chaîne Youtube

4 Analyse en Composantes Principales

Dans l'optique de réaliser une ACP, nous avons extrait un sous-ensemble de notre dataframe. Ce sous-ensemble est uniquement composé, de fait, des variables quantitatives du dataframe initial.

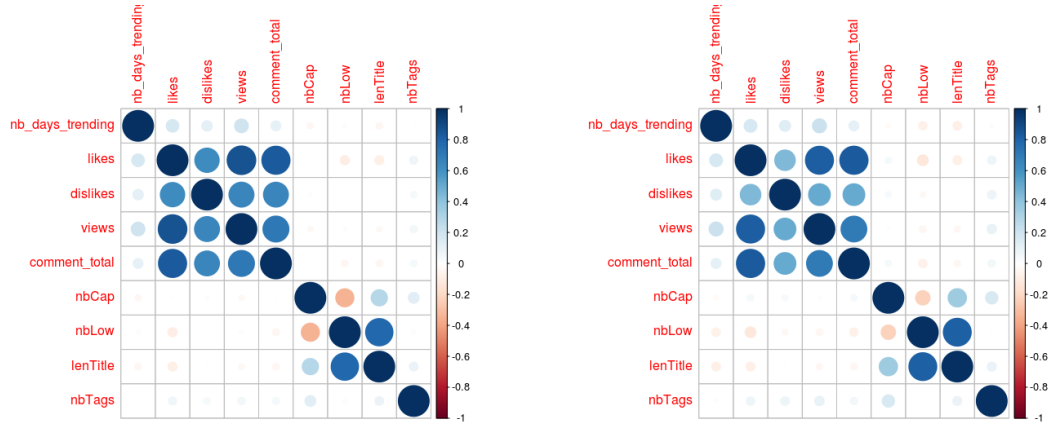
4.1 Nuages de points et matrice de corrélation

Avant de réaliser une ACP sur nos données, nous avons décidé de générer une matrice de corrélation, basée sur les coefficients de corrélation de Pearson. Nos données ne suivant pas de loi normale de manière générale, les coefficients de corrélation calculés ne décrivent que partiellement les associations entre les différentes variables.

```
> shapiro.test(videos_GB.quanti)

        Shapiro-Wilk normality test

data:  videos_GB$views[1:5000]
W = 0.32513, p-value < 2.2e-16

> corrplot(cor(videos_GB[, c("nb_days_trending", "likes",
        "dislikes", "views", "comment_total", "nbCap",
        "nbLow", "lenTitle")]))
```

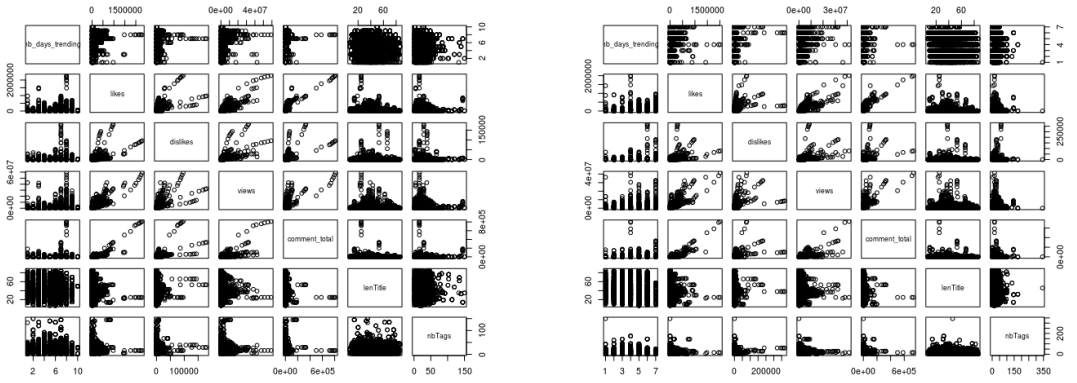
La matrice obtenue nous présente des résultats logiques : le nombre de vues, de likes / dislikes et de commentaires totaux sur les vidéos sont fortement corrélés ($r > 0.5$). Il est également notable que le nombre de majuscules est très légèrement corrélé à la taille du titre ; cela est assez étonnant du fait que le nombre de majuscules est directement extrait du titre.

De même, nous avons souhaité représenter les données sous forme de nuages de points pour tenter d'observer certaines tendances. Les résultats obtenus ont confortés les liaisons mises en évidences par la matrice



(a) GB (b) US

FIGURE 4 – Matrices de corrélation



(a) GB (b) US

FIGURE 5 – Nuage de points présentant la répartition des variables deux à deux

présentée ci-dessus : les variables $likes$, $dislikes$, $views$ et $comment_total$ semblent liées. De même, on peut observer que

Ce graphique nous montre également deux relations intéressantes : il semble que les titres de taille moyennes entraînent un taux de visionnage plus important et qu'un nombre faible de tags soit lié à un plus grand nombre de vues.

```
> pairs(videos_GB.quanti)
> pairs(videos_US.quanti)
```

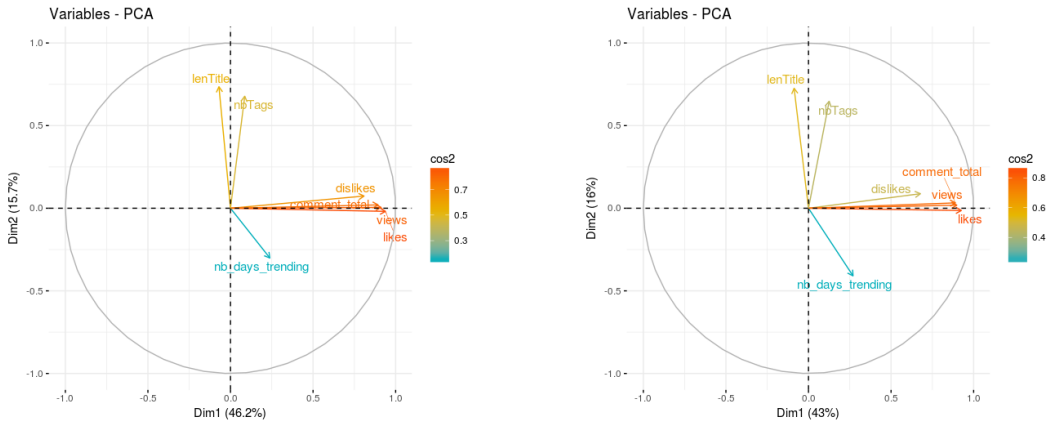
4.2 Réalisation de l'ACP

Pour réaliser notre ACP, nous avons décidé d'utiliser le package **FactorMineR**. Ainsi, à partir des commandes suivantes, nous obtenons le cercle de corrélation résultant de l'ACP ainsi que la contribution de chaque variable aux différentes dimensions.

```
> res.pca <- PCA(videos_GB.active, scale.unit = TRUE)
> var <- get_pca_var(res.pca)
> fviz_pca_var(res.pca, col.var = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800",
              "#FC4E07"),
              repel = TRUE
)
> corrplot(var$contrib, is.corr=FALSE)
```

Les deux dimensions formant ce cercle de corrélation décrivent à 70% notre ensemble de variables. Pour pouvoir décrire celles-ci de manière plus précises - jusqu'à dépasser 80%, en somme - il faut s'intéresser aux 3 premières dimensions. On peut observer que la première dimension sert majoritairement à décrire les variables $likes$, $dislikes$, $comment\_total$ et $views$ tandis que le second sert majoritairement à décrire $lenTitle$ et $nb\_days\_trending$. De même, la 3ème dimension semble également décrire majoritairement ces deux dernières variables.

Ainsi, nous pouvons décrire avec plus de 80% d'efficacité nos 7 variables initiales à partir de 3 dimensions. Ces axes mettent en avant une relation non établie jusqu'à maintenant : celle entre $lenTitle$ et $nb\_days\_trending$.

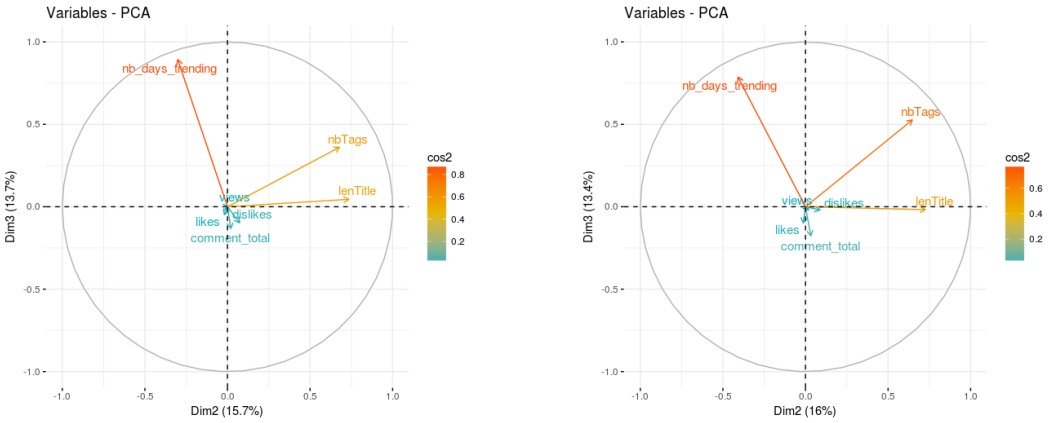


(a) GB (b) US

FIGURE 6 – Cercles de corrélations - Axes 1 & 2

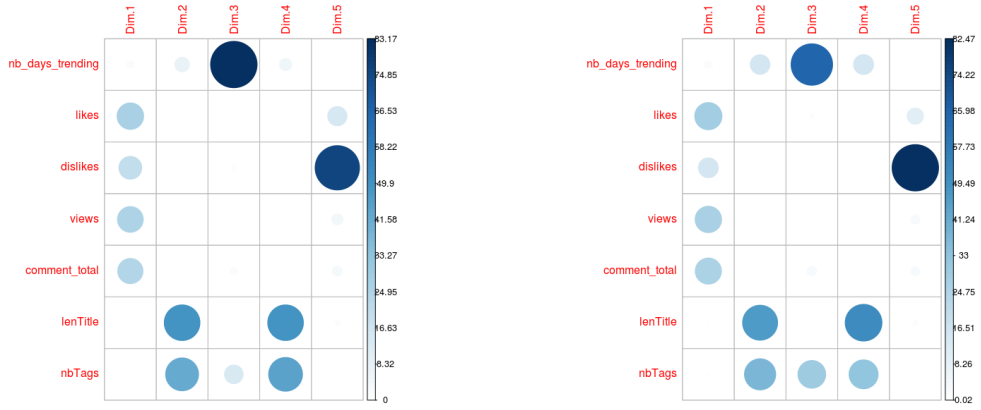
4.3 Interprétation des résultats de l'ACP

Une analyse en Composantes Principales consiste à réduire la dimensionnalité d'un problème, tout en gardant une qualité descriptive intéressante. Dans notre cas, nous avons considéré que le seuil de 80% était suffisant.



(a) GB (b) US

FIGURE 7 – Cercles de corrélations - Axes 2 & 3



(a) GB (b) US

FIGURE 8 – Contribution des variables à chaque dimension

L'ACP réalisée nous a permis de condenser nos 7 données initiales en 3 dimensions. L'influence de chacune de nos variables initiales sur les dimensions trouvées est présentée ci-dessous :

Les liaisons que nous avions proposées auparavant se vérifient ici : la première dimension synthétise les variables $likes$, $dislikes$, $views$ et $comment\_total$. Il est également bon de noter qu'une dimension est - quasiment - réservée à la variable $nb\_days\_trending$.

De par sa nature, l'ACP nous a obligé à considérer uniquement les données quantitatives à notre disposition. Néanmoins, la présence de catégorie et de tags, en plus des titres des vidéos et des chaînes les ayant produit, peut être intéressante à analyser. C'est dans cet objectif que nous nous sommes tournés vers l'Analyse Factorielle Multiple.

5 Analyse Factorielle Multiple

Contrairement à l'ACP, l'AFM n'est pas limitée aux seules données quantitatives. En effet, l'intérêt de celle-ci est de combiner l'analyse en composantes principales et l'analyse des correspondances multiples. L'objectif de cette analyse est, ici, de mettre en avant les effets de groupe des données de même nature. De ce fait, avant de commencer cette analyse, il est nécessaire de classer nos données en groupes et de les normaliser.

5.1 Pré-traitement des données

Comme expliqué ci-dessus, l'analyse factorielle multiple que nous souhaitons effectuer nécessite l'aggrégation de plusieurs données entre elles. Cette aggrégation peut avoir plusieurs natures : on peut décider de les regrouper par type et donc d'avoir deux classes, l'une pour les variables quantitatives et l'autre pour celles qualitatives, ou de les regrouper selon ce qu'elles représentent. La première option fonctionne dans certains cas, majoritairement lorsque les groupes ne sont a priori pas connus. Étant donné la nature de nos données, nous allons ici regrouper 10 de nos variables en 5 groupes qui sont les suivants :

Type de vidéo	Statistiques	Titre	Tags
Channel title, catégorie	Likes, dislikes, views, total_comment	nbCap, nbLow, lenTitle	nbTags

Nous avons ici volontairement exclu plusieurs variables, à savoir $title$ et $tags$ du fait de leurs nombres très importants de modalités.De même, nous avons décidé de conserver la variable $date$ en variable supplémentaires. Une fois ce regroupement effectué, nous avons normalisé les variables quantitatives à notre disposition.

```
> videos_GB_reordered <- videos_GB[, c(2,3,4,6,7,8,9,12,13,14,15,16)]
> videos_GB_ordered$col <- scale(videos_GB_ordered$col)
```

5.2 Interprétation de l'Analyse Factorielle Multiple

De même que pour l'ACP, l'AFM tente de représenter les variables - ou ici les groupes de variables - lui étant présentées selon un certain nombre de dimensions. L'intérêt de ce changement de représentation est la réduction de dimension et la mise en évidence des intéractions entre les groupes introduits.

Dans notre cas, la réduction de dimension n'a pas fonctionné du tout. En effet, l'inertie totale avec 10 dimensions avoisine les 10%. Malgré la refonte des groupes créés précédemment, les résultats obtenus ne se sont pas améliorés.

Dès lors, n'ayant pu expliquer la relation de groupe via l'AFM, nous avons décidé d'utiliser une méthode de clustering pour approfondir cet aspect de nos données.

6 Classification Hiérarchique Ascendante

La classification hiérarchique ascendante est une méthode de classification itérative somme toute simple : on cherche à minimiser la dissimilarité entre les différents individus disponibles, puis on les regroupe jusqu'à ne plus en avoir à disposition. Néanmoins, le dendogramme qu'elle permet d'obtenir est l'une des représentations graphiques de clustering les plus efficaces du fait de son interprétabilité aisée.

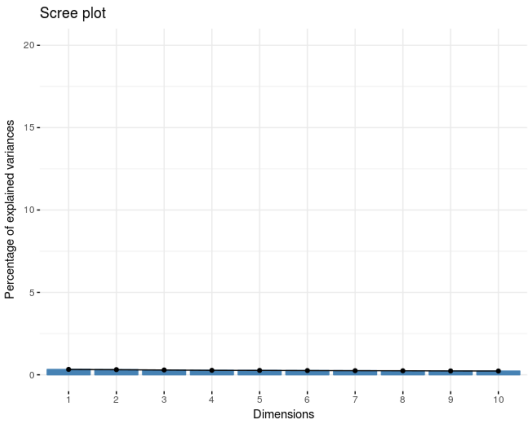


FIGURE 9 – Inertie de chaque dimension générée par l'AFM

6.1 Génération du dendogramme

Comme précédemment, nous avons écarté plusieurs variables. $date$, par exemple, n'a pas été ici considéré de par sa faible relation avec les vidéos en tant que tel. Si ces dates avaient été associées au jour de la semaine, nous aurions peut-être pu montrer un relation hebdomadaire entre les vidéos, mais ce n'est pas le cas ici.

```
> dist.GB <- dist(videos_GB[-c(1,2)])
> cah.ward <- hclust(dist.GB, method="ward.D2")
> plot(cah.ward, axes=FALSE)
> group.cah <- cutree(cah.ward, k=15)
> rect.hclust(cah.ward, k=15)
```

Étant donné le nombre de vidéos présentes dans la base, il était prévisible que le dendogramme généré par cette méthode soit peu lisible. Néanmoins, en analysant les classes formées par cette méthodes et en réduisant le nombre de clusters à 15 (le nombre de catégorie disponibles), nous avons pu remarquer que les vidéos membres d'un même cluster avait généralement la même catégorie !

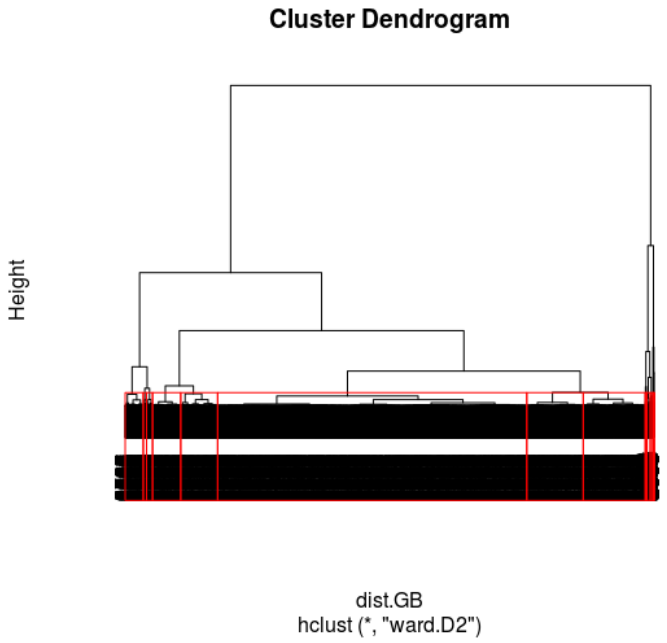


FIGURE 10 – Dendogramme résultant de la CAH

7 Conclusion

La base de données que nous avons choisi d'utiliser comportait initialement peu de données. Néanmoins, le thème de ce dataset étant original, nous avons décidé d'y ajouter des données tout en limitant le biais introduit de ce fait. Les différentes analyses menées ont permis de mettre en évidence plusieurs relations entre les différentes variables quantitatives à notre disposition. L'analyse couplée avec les variables qualitatives n'a malheureusement pas fourni de résultats exploitables. Enfin, la dernière classification mise en oeuvre a réussi à montrer un lien entre l'ensemble de nos variables et les catégories des vidéos étudiées.

Il nous semblerait ainsi intéressant de tenter de réaliser des modèles prédictifs permettant de prédire la catégorie d'une vidéo à partir des variables étant à notre disposition.