:— title: “Projet d’Analyse et de Fouille de Données Massives” output: html_document: df_print: paged pdf_document: default —

Projet d’Analyse et de Fouille de Données Massive

Introduction :

Le but du projet est de manipuler différentes méthodes d’analyses de données sur un jeu de données constitué d’une centaine d’individus et d’une dizaine de variables qualitatives ou quantitatives. Nous avons choisi comme jeu de données les vidéos présentes dans la catégorie “Tendances” du site YouTube France sur une période allant de novembre 2017 à juin 2018. La catégorie “Tendances” de YouTube met en évidence les vidéos les plus vues et les plus appréciées par les utilisateurs de la plateforme. Elle permet donc une plus grande visibilité pour les créateurs de contenu. Une vidéo est placée dans la catégorie “Tendances” par un algorithme se fondant sur les statistiques de la vidéo ainsi que sur les interactions des utilisateurs avec celle-ci. L’analyse des facteurs permettant à une vidéo d’être présente en “Tendances” peut donc être intéressante pour les vidéastes de la plateforme souhaitant être mis en avant.

Présentation de nos données :

Nous avons extrait nos données du site Kaggle (https://www.kaggle.com/datasnaek/youtube-new (https://www.kaggle.com/datasnaek/youtube-new)). Ces données contiennent un ensemble de statistiques sur les vidéos de la catégorie “Tendances” de YouTube. Nous avons décidé d’extraire les statistiques qui nous paraissaient pertinentes pour pouvoir déterminer les facteurs influant sur la présence ou non d’une vidéo dans la catégorie “Tendances”. Ces statistiques sont resumées dans le tableau suivant et seront utilisées comme variables qualitatives ou quantitatives dans les différentes méthodes d’analyse des données.

Nom	Attribut
Taux LOWER Case	quantitatif
Taux UPPER Case	quantitatif
Nb tags	quantitatif
Nb vues	quantitatif
Nb commentaires	quantitatif
Nb likes	quantitatif
Nb dislikes	quantitatif
Taille de la description (en mots)	quantitatif
Catégorie	qualitatif

Nom	Attribut
Jour de la semaine	qualitatif
Moment de la journée	qualitatif
Nombre de lien en description	quantitatif

Nous pouvons nous attendre à une forte corrélation entre la présence d'une vidéo en "Tendances" et son nombre de vues, de mentions "J'aime" et de commentaires. Les vidéos sont placées dans la catégorie "Tendances" principalement à cause du fait que les utilisateurs ont beaucoup interragi avec. Néanmoins, d'autres variables peuvent influer de manière indirecte sur les paramètres cités précedemment.

Le nombre de majuscules nous semble pertinent, car nous avons remarqué une tendance des vidéastes à utiliser des titres uniquement composés de majuscules afin d'attirer l'oeil des utilisateurs de YouTube, et donc d'augmenter plus facilement le nombre de "vues" sur la vidéo.

Le nombre de tags est un autre paramètre important car YouTube propose des recommandations personnalisées aux utilisateurs en fonction de leurs centres d'intérêts. Une vidéo pourrait donc bénéficier d'une audience plus importante si elle était recommandée à plusieurs publics différents. De plus, certains tags sont considérés comme "tendance" lorsqu'ils sont en lien avec l'actualité. Ainsi, lors des incidents récents entre l'Iran et les Etats-Unis, le tag "WWIII" est devenu fréquent sur la plateforme car de nombreux internautes craignaient une détérioration de la situation. Cela a permis à certains créateurs de contenu politique, par exemple, d'exploiter cette tendance en publiant du contenu connexe à ce tag.

Un autre élément qui pourrait nous donner des indicateurs sur la mise en avant d'une vidéo par l'algorithme de YouTube serait l'espace de description de la vidéo. Chaque vidéo doit posséder une description. Les descriptions sont généralement utilisées pour résumer la vidéo, mettre à disposition des utilisateurs les liens vers les sources de la vidéo, les différentes pages sur les réseaux sociaux du créateur mais aussi vers des magasins en ligne, des sites de sponsors ou alors d'autres chaînes YouTube dans le cadre d'une collaboration. Nous allons plus particulièrement nous intéresser à la longueur de la description. En effet, la majorité des créateurs qui vivent de leur activité sur YouTube possèdent des descriptions longues et détaillées. Cela peut bien sûr aussi être le cas pour des vidéastes plus amateurs. Nous avons également compté le nombre de liens que possédait la description. En effet, mis à part les chaînes d'information et de vulgarisation, la majorité des chaînes proposent un lien vers un magasin en ligne, trois ou quatre liens vers des pages de réseaux sociaux et éventuellement des liens vers des sites de sponsors. Cependant, lorsqu'une vidéo est le fruit d'une collaboration avec d'autres créateurs, un lien vers la chaîne YouTube de cette autre personne est généralement présent dans la description. Les vidéos réalisées avec plusieurs vidéastes attirent un public plus large et sont devenues beaucoup plus communes qu'auparavant sur la plateforme. Nous esperons donc, à travers cette variable, mettre en avant ce phénomène.

La publication d'une vidéo peut également être ciblée afin de toucher le plus de personnes possible (une vidéo publiée le vendredi soir à 18h aura probablement plus d'impact qu'une vidéo publiée le mardi matin à 8h). Nous avons donc choisi de nous intéresser au moment auquel une vidéo a été mise en ligne. Nous avons découpé de la manière suivante les moments de la journée pour nos analyses : 00:00 -> 11:59 = "morning", 12:00 -> 19:00 = "afternoon" et le reste est

considéré comme “evening”.

Nous serons également en mesure d’exhiber les types de vidéos qui sont le plus visionnées par le public français, à travers la catégorie de la vidéo. Ce facteur nous permettra également de comparer les audiences des vidéos appartenant à différentes catégories (par exemple, les personnes regardant des vidéos sportives seraient peut-être plus enclines à consulter des vidéos de mode ou de divertissement que les personnes visionnant des vidéos “Gaming”).

Traitement du jeu de données

A partir du jeu de données télechargé au format csv, nous avons généré un nouveau fichier csv contenant, pour chaque vidéo, les statistiques listées dans le tableau précédent. Vous pouvez retrouver le code commenté pour l’extraction et le traitement des données brutes initiales dans le fichier ‘processData.py’.

Analyse des données

Dans cette partie, nous allons analyser les données que nous avons extraites et qui se trouvent dans le fichier “youtubeTrends.csv”. Ce fichier contient 40703 individus. Néanmoins les méthodes d’analyse détaillées dans cette partie ne porteront que sur une centaine d’invidus. Nous sélectionnons donc 100 vidéos de manière aléatoire dans le fichier “youtubeTrends.csv”.

Code pour selectionner les 100 individus aléatoirement :

```
# data = read.csv("./DATA/youtubeTrends.csv")
# data = data[sample(nrow(data), 100),]
```

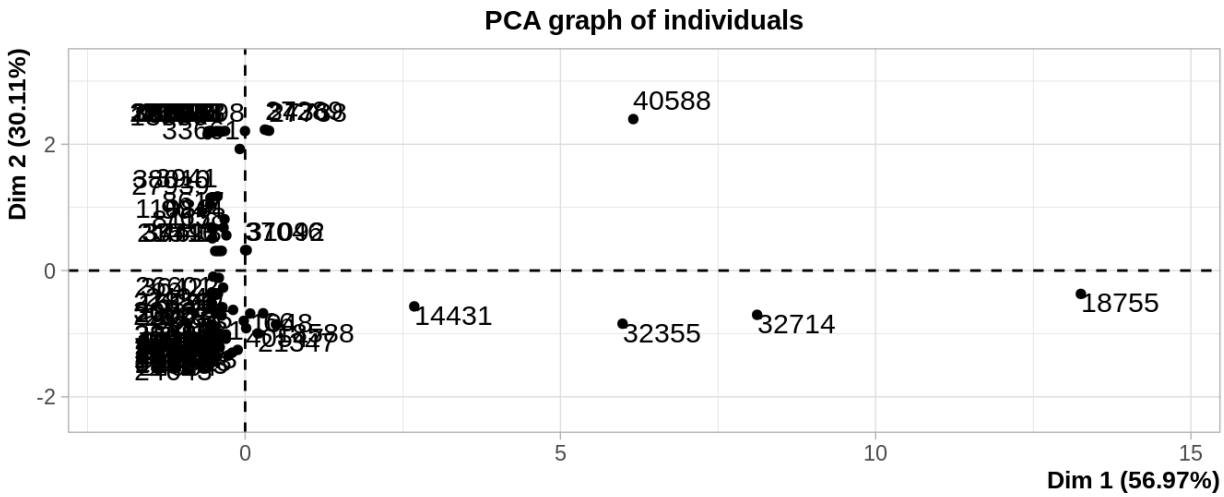
Code utilisé pour récupérer les 100 individus :

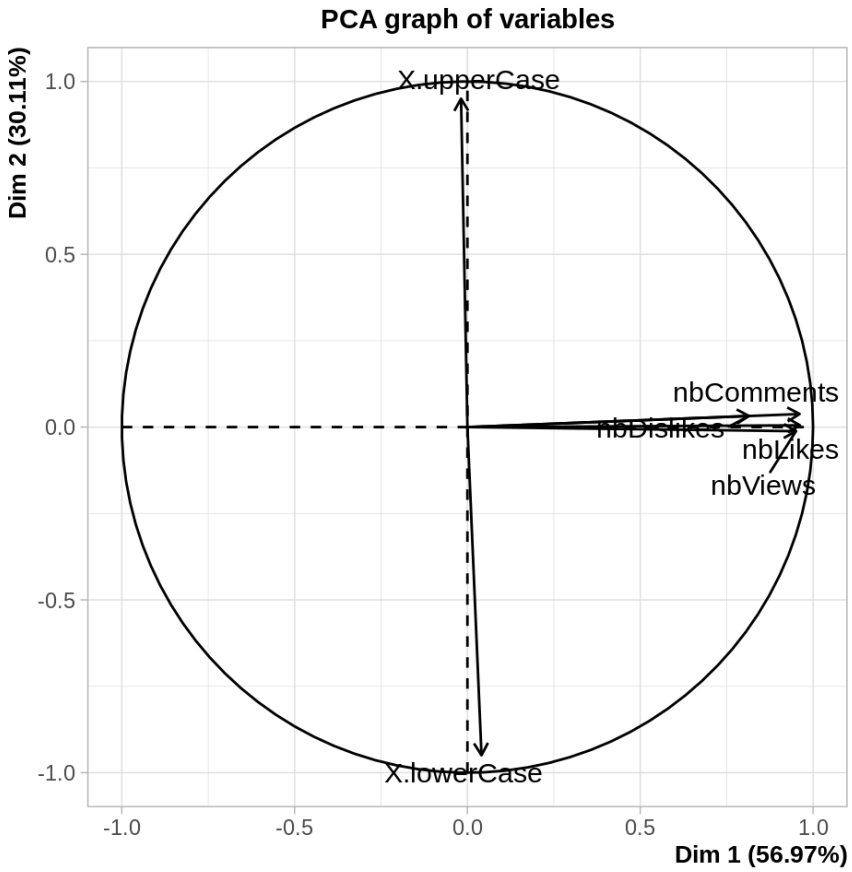
```
load('./random_data.rdata')
```

Analyse en Composantes Principales (ACP)

La première analyse que nous allons effectuer sera une ACP. Cette analyse ne porte que sur des données qualitatives. Ainsi, il faut retirer de nos données les variables qualitatives. Ensuite, on applique l’ACP sur les données filtrées.

```
# On retire les données qualitatives
data_ACP = data
data_ACP$category <- NULL
data_ACP$momentOfDay <- NULL
data_ACP$day <- NULL
data_ACP$index <- NULL

# On retire les données considérées comme faiblement corellées après les pre
miers tests
data_ACP$nbTags <- NULL
data_ACP$nbWords <- NULL
data_ACP$nbLinks <- NULL

# On applique l'ACP sur les données filtrées
res.pca = PCA(data_ACP, scale.unit=TRUE, ncp=6, graph=T)
```

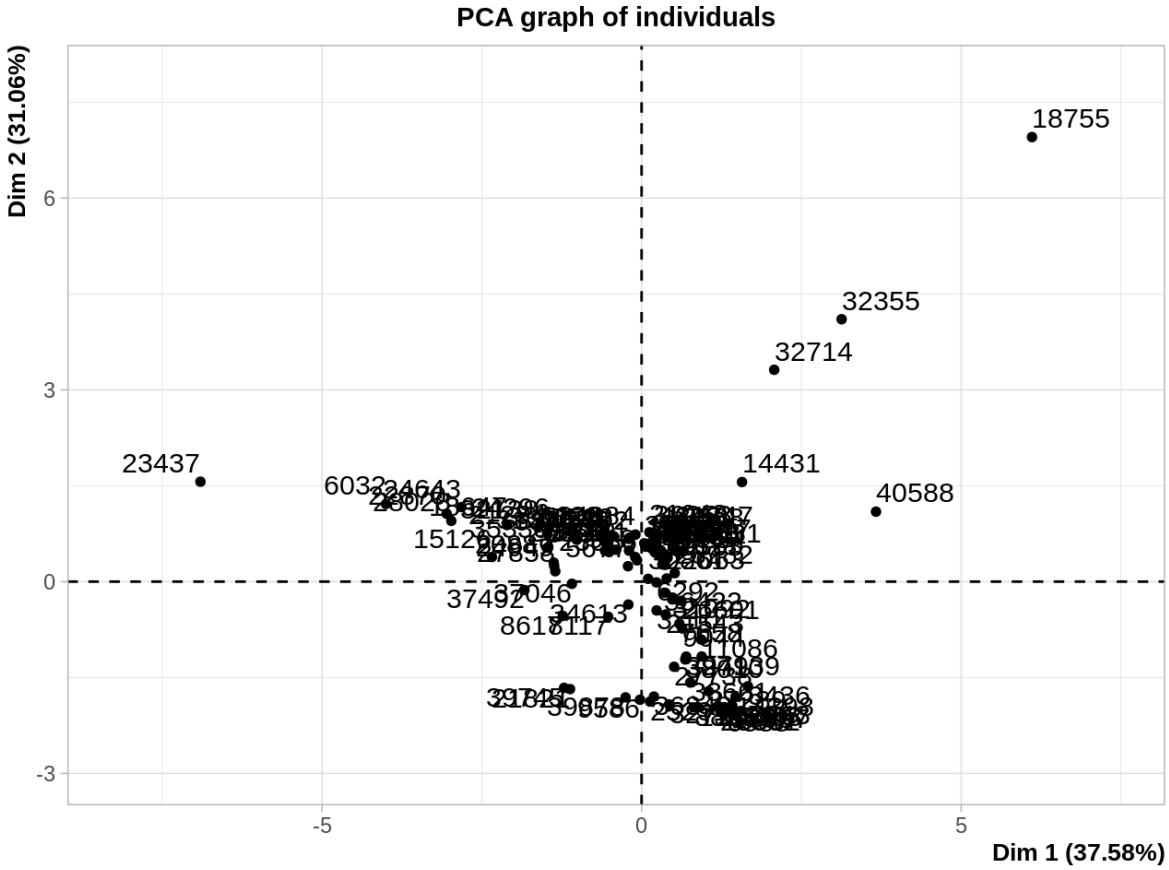


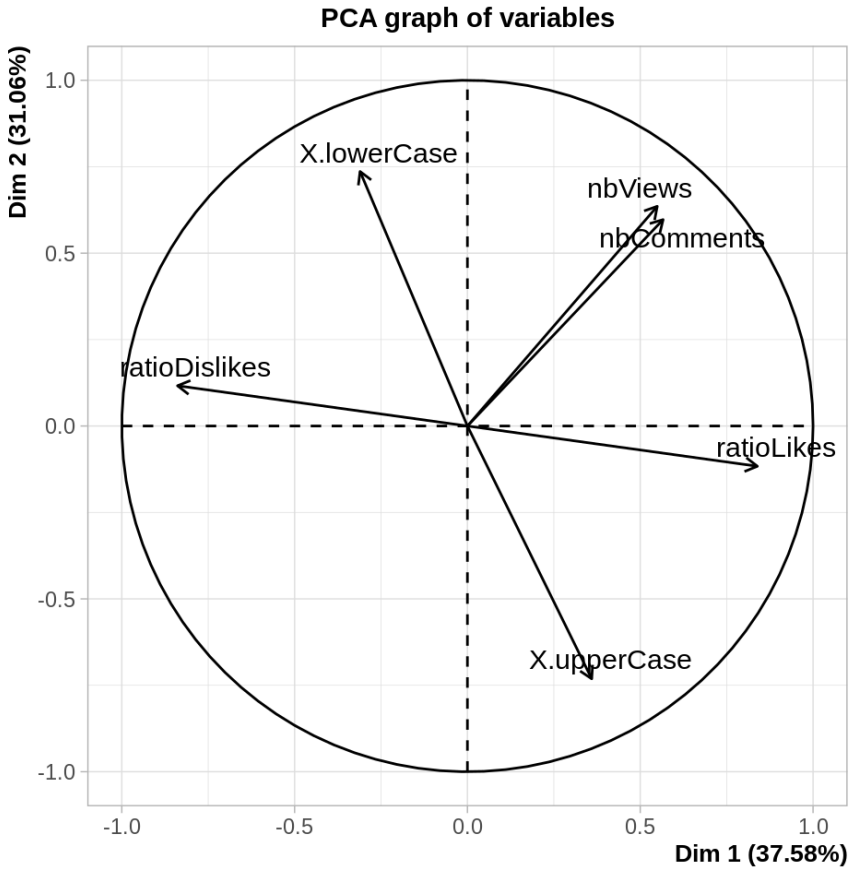


Sur la figure, on constate que les données sont bien représentées. En effet, 46.89% des données sont représentées sur l'axe 1 et 28.47% sur l'axe 2. On voit que le nombre de likes et dislikes sont corrélés. Néanmoins, après analyse, il apparaît que ces deux variables ne capturent pas vraiment ce que l'on voulait. En effet, nous ne savons pas en quoi elles sont corrélées (si un grand nombre de likes est associé à un grand nombre de dislikes…). Ainsi, pour voir l'impact du nombre de likes et de dislikes, on introduit deux nouvelles variables quantitatives : le ratio de like et le ratio de dislikes par vidéo.

```
# On retire les données qualitatives
data_ACP = data
data_ACP$category <- NULL
data_ACP$momentOfDay <- NULL
data_ACP$day <- NULL
data_ACP$index <- NULL

# On retire les données faiblement corellé
data_ACP$nbTags <- NULL
data_ACP$nbWords <- NULL
data_ACP$nbLinks <- NULL

# Ajout du ratio de like et dislike
data_ACP$ratioLikes <- data_ACP$nbLikes/(data_ACP$nbLikes + data_ACP$nbDisli
kes)
data_ACP$ratioDislikes <- data_ACP$nbDislikes/(data_ACP$nbLikes + data_ACP$n
bDislikes)

data_ACP$nbLikes <- NULL
data_ACP$nbDislikes <- NULL

res.pca = PCA(data_ACP, scale.unit=TRUE, ncp=6, graph=T)
```

```
## Warning in PCA(data_ACP, scale.unit = TRUE, ncp = 6, graph = T): Missing
values
## are imputed by the mean of the variable: you should use the imputePCA fun
ction
## of the missMDA package
```





On observe que le ratio de likes et de dislikes sont inversement corrélés. En effet, cela semble cohérent qu'une vidéo en "Tendances" ayant un fort nombre de mentions "J'aime" ait une faible proportion de mentions "Je n'aime pas" et c'est bien ce que l'on constate sur le graphe résultant de

l'ACP.

Analyse des Composantes Multiples (ACM)

Dans cette partie, nous allons réaliser une ACM sur 3 attributs qualitatifs : le jour de la semaine où la vidéo est publiée, le moment de la journée et la catégorie de la vidéo.

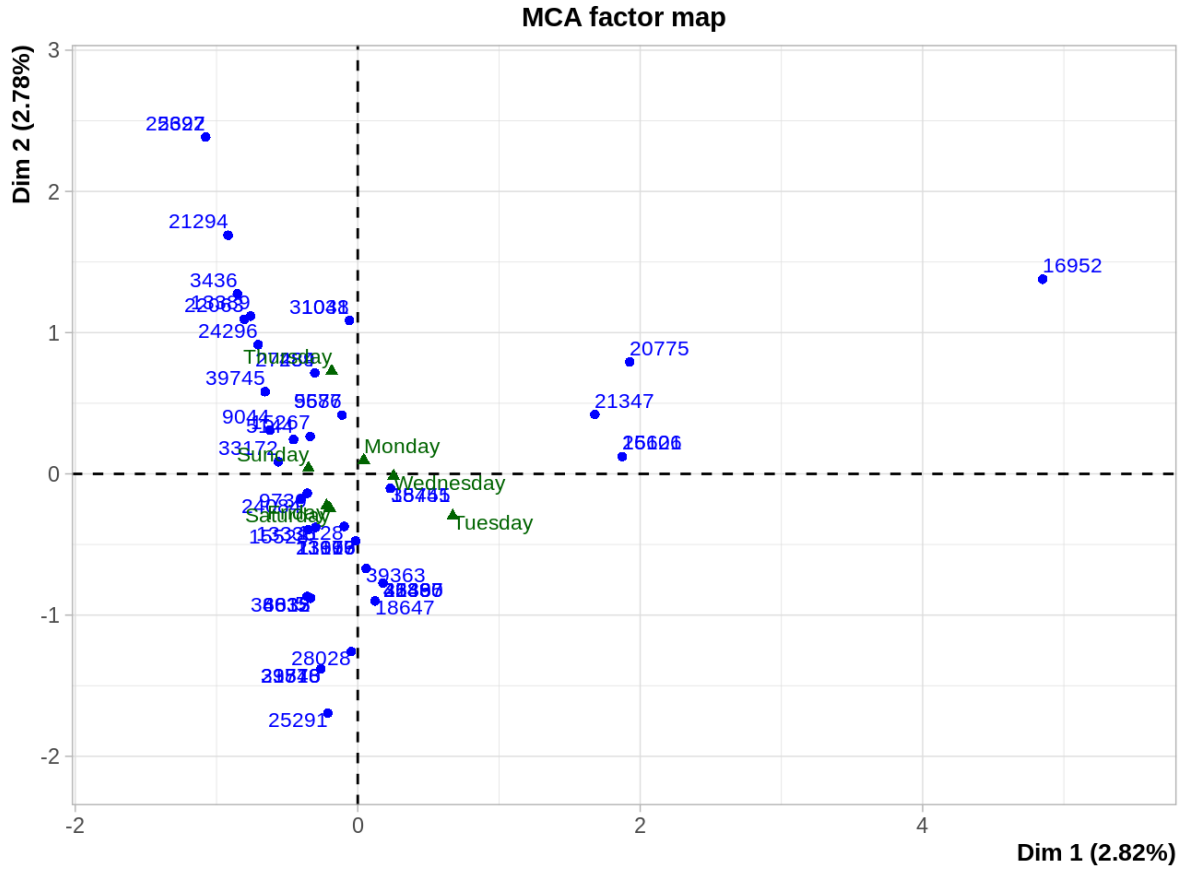
Dans un premier temps, nous préparons les données en enlevant la moitié des données. En effet, si l'on choisit d'effectuer l'ACP sur l'ensemble des données, le graphe résultant contient trop de points superposés et il est difficile de l'interpréter. Ensuite, il faut effectuer une petite opération pour transformer nos données en type Factor (un type spécifique à R). Cette opération est necessaire car sans cela, nous obtenons une erreur. Enfin, on retire tous les attributs quantitatifs sauf le nombre de likes, dislikes et de commentaires. On conserve ces attributs car on souhaite regrouper les attributs qualitatifs en fonction de leur poids sur le fait que la vidéo soit en "Tendances", qui se représente à l'aide des 3 variables quantitatives ci-dessus.

```
data_ACM = data[1:50,]
i=0
while(i<ncol(data_ACM)){
  i=i+1
  data_ACM[,i]=as.factor(data_ACM[,i])
}
data_ACM$index <- NULL
data_ACM$nbTags <- NULL
data_ACM$nbWords <- NULL
data_ACM$nbLinks <- NULL
data_ACM$X.lowerCase <- NULL
data_ACM$X.upperCase <- NULL
```

Moment de la semaine

Nous nous interessons, dans un premier temps, au moment de la journée où la vidéo à été publiée. On s'attend éventuellement à voir regroupés ensemble des jours de la semaine qui sont proches (en fonction de s'ils sont en début, milieu et fin de semaine).

```
res.mca_1 = MCA(data_ACM, quali.sup = c(4), graph = FALSE)
plot.MCA(res.mca_1, invisible=c("var"), cex=0.75)
```

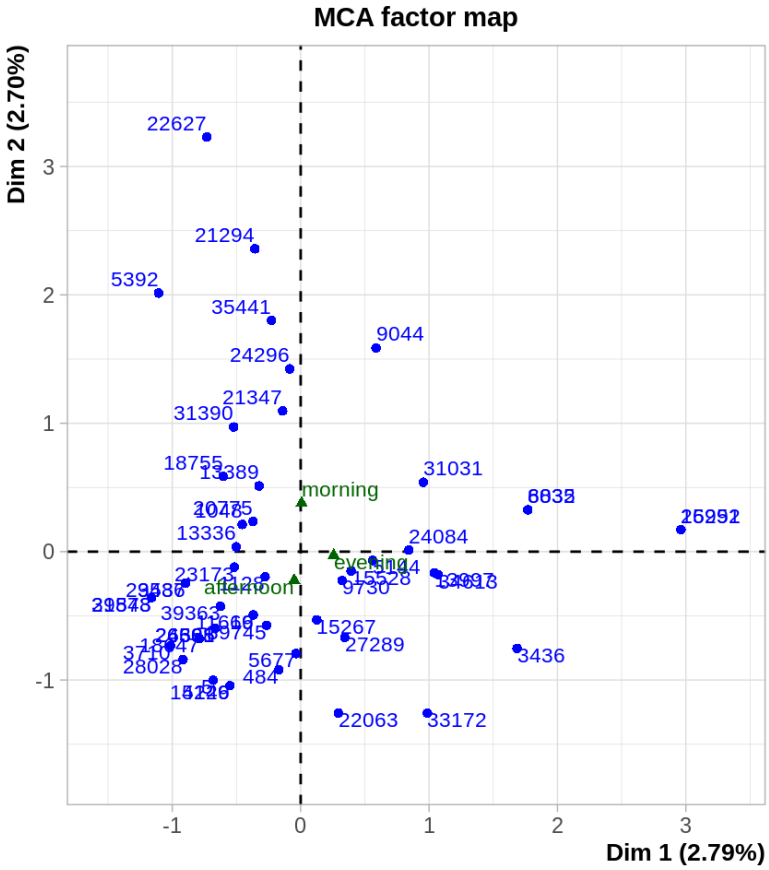


Les résultats obtenus sont plutôt satisfaisants, même si l'on représente peu les données (2.82%). On peut observer que certains jours sont reliés comme le Lundi, Mardi et Mercredi. Cela semble cohérent qu'ils aient la même visibilité, car ils se trouvent tous trois en début de semaine. De la même manière, le Vendredi est extrèmement corrélé au Samedi et au Dimanche. Résultat plus étonnant, on constate que le Jeudi est isolé dans le graphique, les vidéos publiées le Jeudi doivent donc être visualisées par un public ciblé.

Moment de la journée

La prochaine ACM cherche à étudier les corrélations entre les moments de la journée où sont publiés une vidéo. Intuitivement, on pourrait penser que les vidéos qui sortent l'après-midi n'ont pas le même "succès" qu'une vidéo qui a été publiée le matin ou le soir.

```
res.mca_2 = MCA(data_ACM, quali.sup = c(5), graph = FALSE)
plot.MCA(res.mca_2, invisible=c("var"), cex=0.75)
```

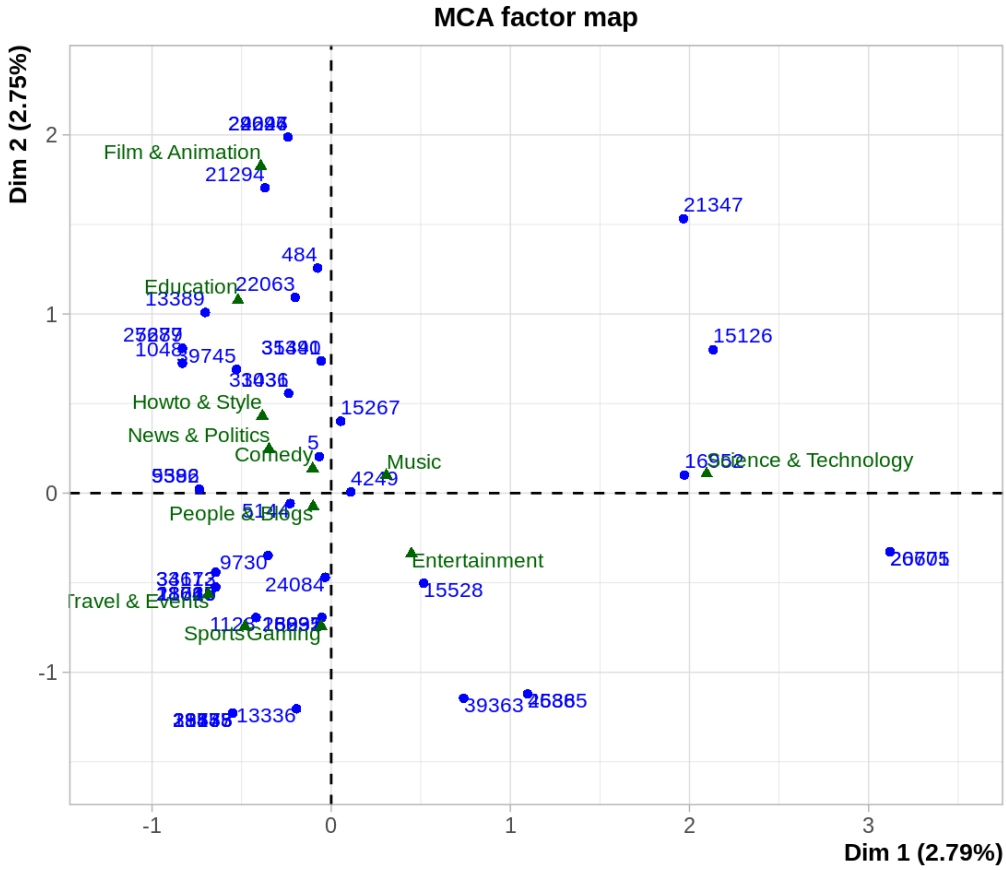


Sur le graphique résultant, on constate que les individus peuvent être regroupés en deux catégories. Tout d'abord, la majorité des vidéos en "Tendances" sont publiées l'après-midi ou le soir. Néanmoins, un certain nombre de vidéos présentes dans la catégorie "Tendances" ont été mises en ligne le matin. Cela peut s'interpréter de deux manières. D'une part, ces vidéos peuvent être destinées à un public particulier, pouvant les regarder le matin. D'autre part, ces vidéos peuvent être brèves (comme celles présentes sur les réseaux sociaux) donc pouvant être visionnées rapidement par les utilisateurs et/ou traitant d'actualité.

Catégorie

Enfin, la dernière ACM porte sur les catégories des vidéos présentes dans la catégorie "Tendances".

```
res.mca_3 = MCA(data_ACM, quali.sup = c(7), graph = FALSE)
plot.MCA(res.mca_3, invisible=c("var"), cex=0.75)
```

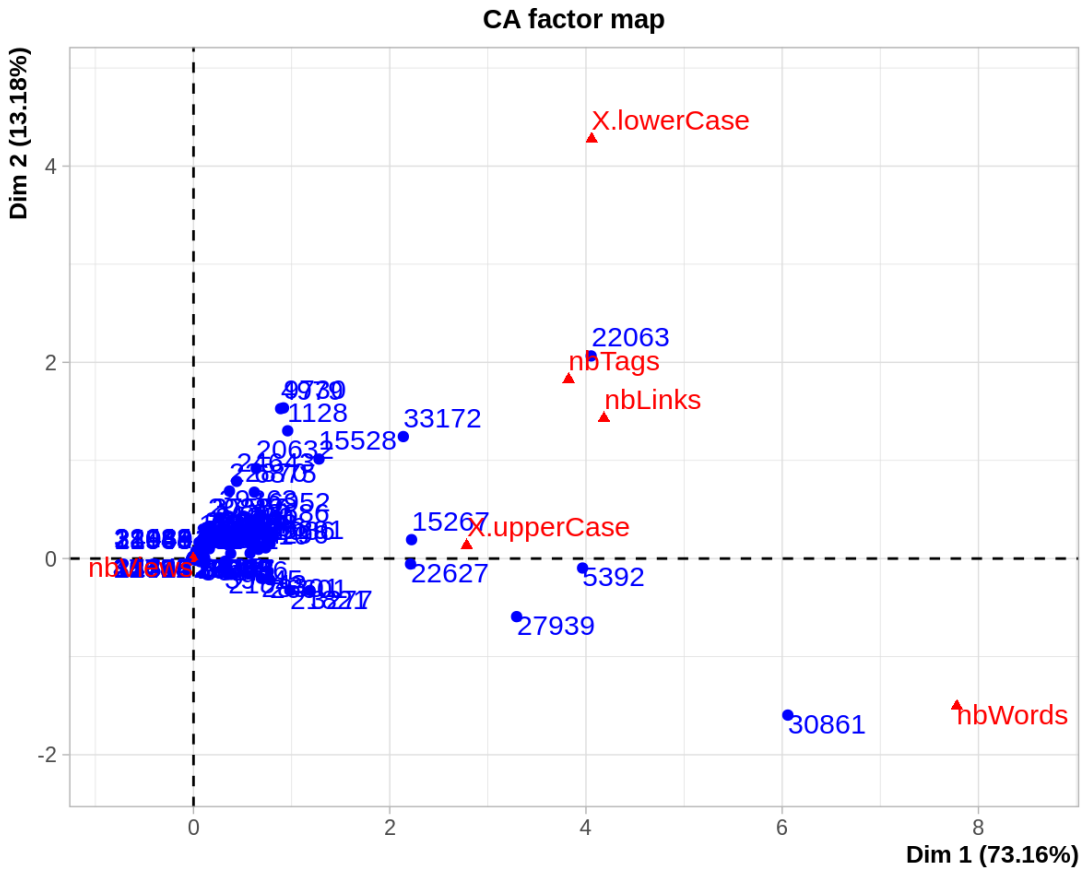


Nous remarquons que de nombreuses catgéories sont liées entre elles dans le sens où elles ont des audiences communes. Ainsi, sur le graphique, nous constatons que les personnes regardant des vidéos de la catégorie “Howto & Style”, visionnent également des vidéos appartenant aux catégories “News and Politics” et “Comedy”. De manière similaire, les personnes regardant des vidéos de la catégorie “Gaming” visionnent également des vidéos traitant de sport et de voyage.

Analyse Factorielle des Correspondances (AFC)

La dernière analyse que nous proposons est l’AFC. Grâce à cette méthode, nous allons mettre en évidence les attributs qui fédèrent le plus d’individus.

```
data_AFC = data
res.afc = CA(data_AFC[, c(2:5 ,12:13)])
```



La figure obtenue est très intéressante. Tout d'abord, on constate que toutes les vidéos en tendance partagent un nombre de vues relativement proche. Le nombre de liens en description de la vidéo est également proche du nuage de points. Cela confirme que les vidéos collaboratives entre créateurs génèrent plus de visionnages. Un autre attribut proche du nombre de liens est le nombre de tags de la vidéo. Les tags aident à mettre la vidéo en avant et sont utiles, notamment, si la vidéo traite de l'actualité. De plus, certains tags peuvent momentanément profiter d'une certaine notoriété. A la sortie d'un nouveau jeu vidéo, il est probable que le tag "Jeux" ou "Gaming" soit mis en avant. Enfin, on peut remarquer que l'attribut "UpperCase" est plus proche du nuage de points que l'attribut "LowerCase". Cela semble indiquer qu'écrire les titres en majuscule attire davantage les utilisateurs du site.

Conclusion

Nous avons analysé les données issues des vidéos de la catégorie "Tendances" du site YouTube (France). Ce travail nous a permis d'appliquer les méthodes d'analyse et de fouille de données que nous avons vues en cours. Nous avons ainsi pu confirmer certaines hypothèses sur les facteurs aboutissant au fait qu'une vidéo soit présente en "Tendances" mais aussi découvrir des liens qui nous ont étonnés.