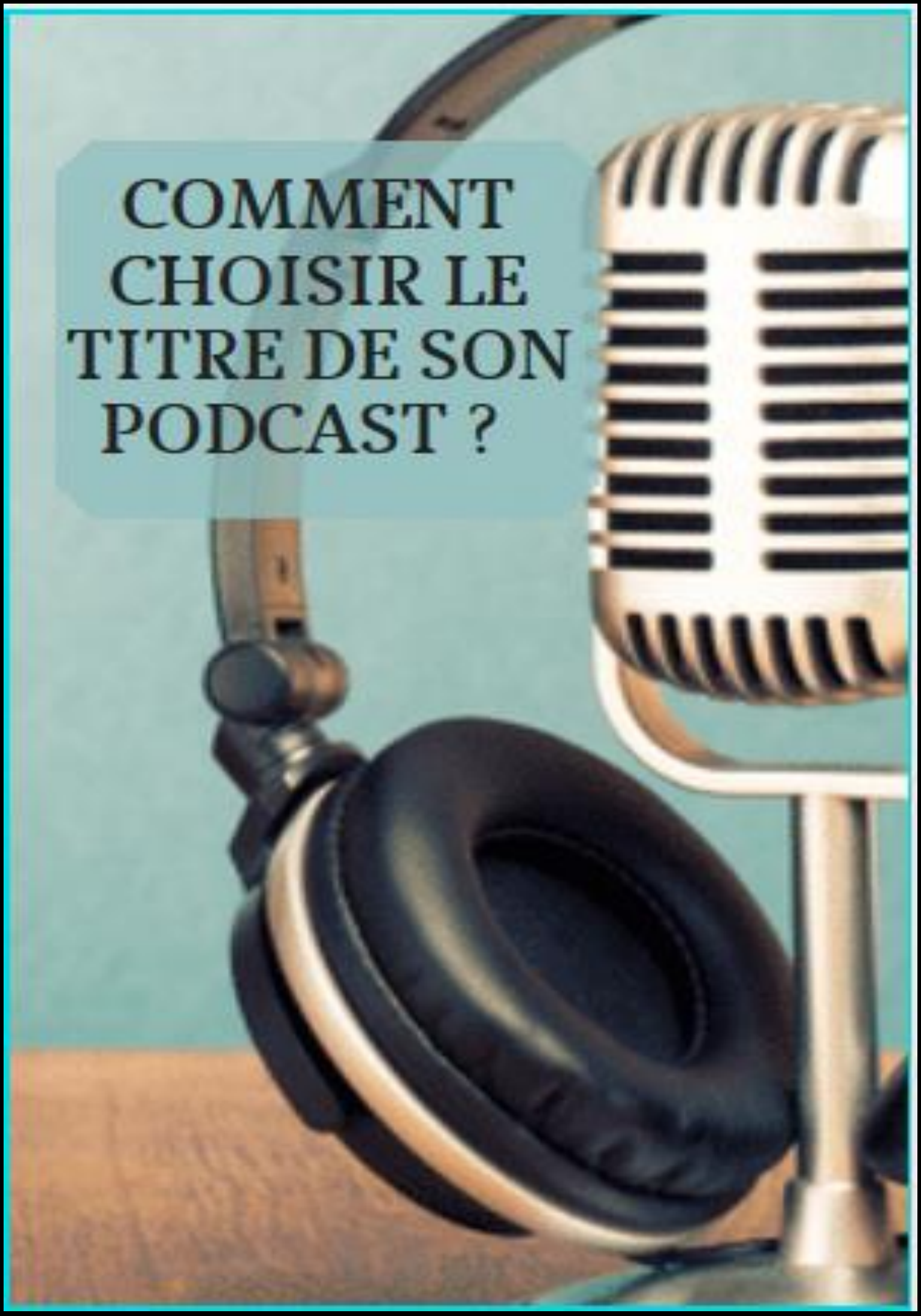

COMMENT
CHOISIR LE
TITRE DE SON
PODCAST ?

Les coordonnées personnelles du groupe

Nom : ***SCHOMBOURGER***

Prénom : ***Hugo***

N° d'étudiant : **11620295**

Adresse mail : ***hugoschombourger@gmail.com***

Numéro de téléphone : **0685494984**

Nom : ***KHELID***

Prénom : **Jalis**

N° d'étudiant : **11908450**

Adresse mail : **jaliskhelid@icloud.com**

Nom : **NAKHAAI**

Prénom : **Sara**

N° d'étudiant : **11910498**

Adresse mail : **sara.nakhaai@gmail.com**

Nom : **BOITI**

Prénom : **Hind**

N° d'étudiant : **11633431**

Adresse mail : **boitihind111@gmail.com**

Nom : **ZABALMOUGAMADOU**

Prénom : **Mohamed Tahsin**

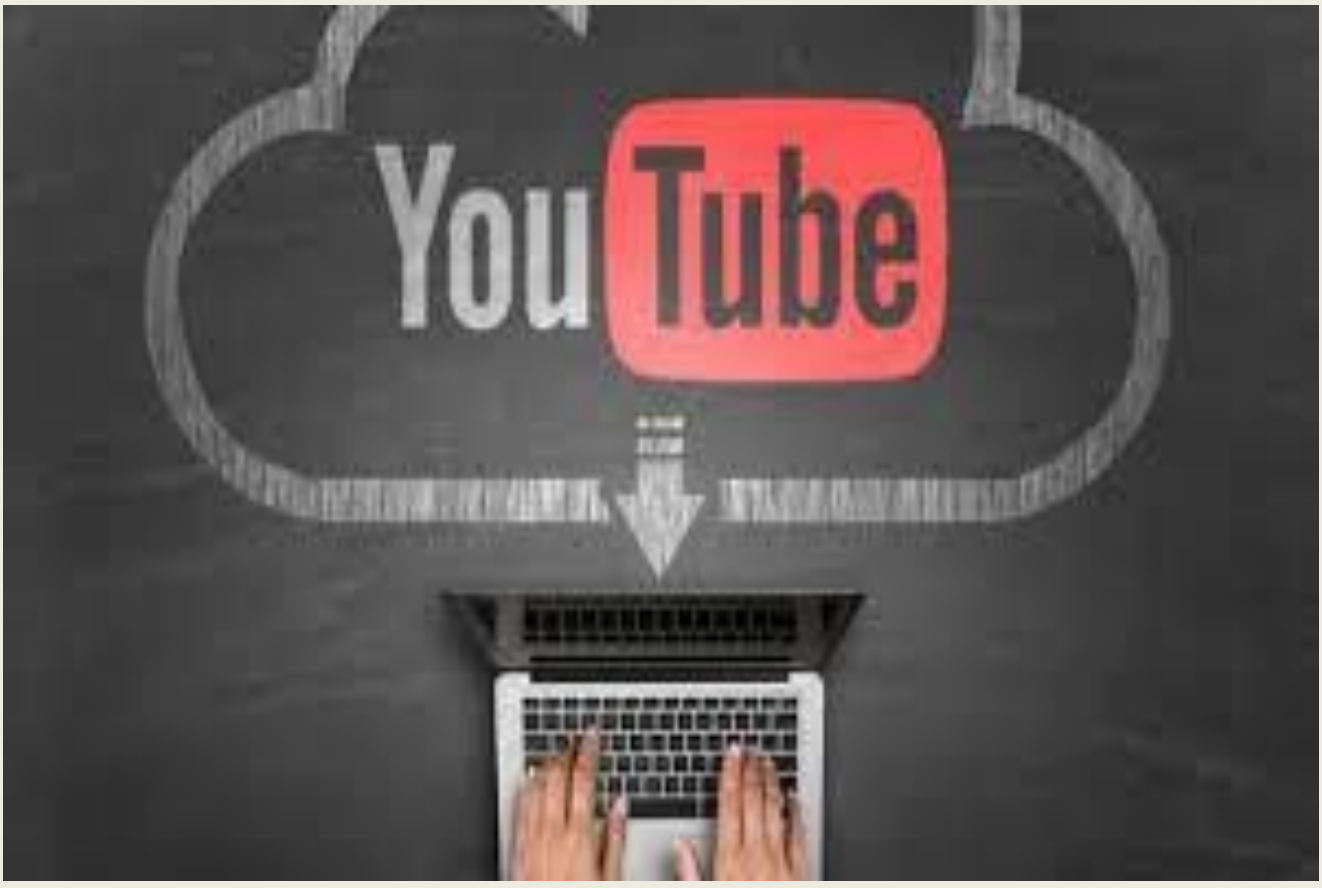
N° d'étudiant : ***11700272***

Adresse mail : ***mohamedtahsin1@gmail.com***

Table des matières :

I. Introduction et objectif de la recherche

A l'origine notre idée était de faire une recherche sur les critères influençant le nombre de vues des vidéos sur YouTube mais cela aurait introduit des biais car le visuel sur la miniature des vidéos influence nécessairement les internautes et mesurer cet aspect est très subjectif et complexe. Nous avons donc décidés de recentrer notre étude sur les podcasts éliminant ainsi les biais introduits par l'aspect visuel, et de faire une recherche sur l'influence des aspects quantitatifs d'un titre(comme le nombre de caractères) sur le nombre de vue des podcasts.



II. Méthodologie :

Collecte de données :

Nous avons collecté les données de cette façon :
Nous avons choisi uniquement des chaînes dans lesquelles le visuel sur la miniature des vidéos sont quasiment identiques afin d'éviter les biais introduits par l'aspect visuel.
Sur un échantillon de 16 chaînes YouTube différentes postant des Podcasts nous avons collectés les données de 160 vidéos (10 par chaîne). Les caractères mesurées étaient les suivants :

- **le nombre de vue du podcast en question,**
- **le "ratio nombre de vue" qui est égal au nombre de vues de la vidéo divisé par la moyenne de nombre de vues des vidéos de la chaîne. Ceci permet d'éliminer l'influence du nombre d'abonnés d'une chaîne sur le nombre de vue. Avec cette harmonisation, deux vidéos de deux chaînes différentes sont maintenant comparables. Ainsi, un ratio de 1.2 pour une vidéo signifie que le nombre de vue de cette vidéo est supérieur de 20% par rapport à la moyenne des vues des vidéos de la chaînes**
- **la durée du podcast en minutes**
- **le nombre de points d'exclamation**
- **le nombre de points d'interrogation**
- **le nombre de caractères spéciaux (exemple:&,#,§…)**
- **le nombre de chiffres**
- **le nombre de majuscules**
- **la longueur du titre mesurée en nombre de caractères**

Les données ont été regroupés dans un fichier Excel, ce qui nous a permis d'obtenir le tableau suivant:

nbrdevue	rationbrvue	dureeminute	nbrinterrogat	nbrexclamati	nbrcaractspe	nbrchiffre	nbrmaj	nbrcaract
1900000	0.84	95	0	0	2	4	5	34
2800000	1.24	85	0	0	2	4	5	36
789000	0.35	116	0	0	2	4	5	36
1600000	0.71	149	0	0	2	4	5	35
1800000	0.79	119	0	0	2	4	5	33

Analyse des données :

Afin d'analyser les données collectés et de faire ressortir les liens entre les variables explicatives et le "ratio nombre de vue" (qui représente dans notre cas le succès d'une vidéo),nous avons utilisé le logiciel R.

L'avantage d'utiliser le "ratio nombre de vue" au lieu du "nombre de vues" est que le ratio élimine l'influence du nombre d'abonnés sur le nombre de vue de la vidéo. Si on n'élimine pas cet effet, alors deux vidéos auront un nombre de vues différent pour un même nombre de caractère, ce qui nous empêche d'étudier l'effet des variables mesurées sur le nombre de vue.
Le "ratio nombre de vues" mesure la variation du nombre de vue par rapport au nombre de vue moyen de la chaîne. Avec cette harmonisation, on élimine l'effet du nombre d'abonnés sur le nombre de vue. Donc dans la suite nous n'étudierons directement le nombre de vue mais nous passerons par le "ratio nombre de vue" pour analyser les effets sur le nombre de vue.

Nous avons commencé par importer notre fichier Excel contenant les données dans R en faisant projet=read.csv2(choose.files(),dec=".") et nous avons utilisé par la suite la fonction GLM "General Linear Model".

Dans un premier temps, nous avons utilisé le GLM sur tout le tableau en négligeant l'effet de chaque variable explicative sur l'autre. Nous avons nommé le résultat "premiermodel".

Ensuite, pour intégrer les interactions entre les différentes variables, nous avons utilisé la puissance 2 (^2) sur les variables ce qui permet d'avoir l'effet de chaque variable seule et l'effet de deux variables combinés. Ce deuxième résultat a été nommé "deuxiememodel".

Par la suite, nous avons utilisé la fonction Step pour simplifier le modèle et enlever tout ce qui n'est pas significatif.

Pour accéder aux résultats, nous avons appliqué la fonction summary sur les deux modèles. On regarde si c'est significatif ou pas (par la présence des étoiles) et regarde si la première colonne (estimate) est positive ou négative pour trouver dans quel sens varie les variables.

La figure suivante représente le code d'analyse sur R:

```
"On importe notre classeur enregistre sous format .CSV"
projet=read.csv2(choose.files(),dec=".")

"On modifie les noms des variables pour une utilisation plus simple"
nbrdevue<-projet$nbrdevue
rationbrvue<-projet$rationbrvue
dureeminute<-projet$dureeminute
nbrinterrogation<-projet$nbrinterrogation
nbrexclamation<-projet$nbrexclamation
nbrcaractspeciaux<-projet$nbrcaractspeciaux
nbrchiffre<-projet$nbrchiffre
nbrmaj<-projet$nbrmaj
nbrcaract<-projet$nbrcaract

"On fait le glm"
premiermodel=glm(rationbrvue~.,data=projet)
premiermodel=step(premiermodel)
summary(premiermodel)

"On fait le glm avec ^2"
deuxiememodel=glm(rationbrvue~(nbrdevue+dureeminute+nbrinterrogation+nbrexclamation+nbrcaractspeciaux+nbrchiffre+nbrmaj+nbrcaract)^2)
deuxiememodel=step(deuxiememodel)
summary(deuxiememodel)
```

III. Résultats de la recherche

Ces fonctions nous ont permis de retrouver les résultats suivants:

- **GLM sans ^2**
  - **L'influence de la durée de la vidéo sur le nombre de vues est significative (une étoile) et la première colonne est positive, donc le "ratio nombre de vue augmente avec l'augmentation de la durée ce qui signifie que le nombre de vue augmente avec la durée;**

  - **La longueur de titre(le nombre de caractères) de la vidéo a une influence plus significative sur le "ratio nombre de vue"(donc le nombre de vues), plus le titre est long plus la vidéo est vue.**

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.07268  -0.34397  -0.06818   0.21354   3.15370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.109e-01  1.521e-01   2.702  0.00761 **
nbrdevue    6.580e-09  4.388e-10  14.994  < 2e-16 ***
dureeminute 2.322e-03  1.080e-03   2.149  0.03311 *
nbrcaract   1.046e-02  3.315e-03   3.156  0.00190 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.374683)

    Null deviance: 146.134  on 167  degrees of freedom
Residual deviance:  61.448  on 164  degrees of freedom
AIC: 317.79

Number of Fisher Scoring iterations: 2
```

- **GLM avec ^2**
  - **La longueur de titre(le nombre de caractères) de la vidéo a une influence encore plus significative sur le “ratio nombre de vue”(donc le nombre de vues). Plus la longueur du titre augmente plus le nombre de vue augmente.**

  - **L’influence de la durée de la vidéo sur le “ratio nombre de vue”(donc le nombre de vues) est maintenant très significative**
    - **Cette influence de la durée de la vidéo diminue lorsque le nombre de caractères augmente et vice-versa.**
    - **Cette influence de la durée de la vidéo diminue lorsque le nombre de chiffres augmente et vice-versa.**

  - **Le nombre de lettres en majuscules contenues dans le titre augmente le “ratio nombre de vue”(donc le nombre de vues) (une étoile)**
    - **Cette influence du nombre de majuscules diminue lorsque le nombre de caractères (la longueur du titre) augmente et vice-versa.**

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.40532  -0.27956  -0.02222   0.21119   2.73692

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.866e-01  2.233e-01  -0.836 0.404480
nbrdevue                     9.831e-08  7.593e-08   1.295 0.197382
dureeminute                  1.514e-02  4.249e-03   3.563 0.000489 ***
nbrcaractspeciaux            6.625e-02  4.941e-02   1.341 0.181962
nbrchiffre                  -6.598e-02  1.448e-01  -0.456 0.649364
nbrmaj                       2.874e-02  1.287e-02   2.233 0.026996 *
nbrcaract                    2.006e-02  5.316e-03   3.772 0.000231 ***
nbrdevue:nbrcaractspeciaux  -5.348e-08  2.919e-08  -1.832 0.068861 .
nbrdevue:nbrmaj             -4.801e-08  2.167e-08  -2.215 0.028235 *
nbrdevue:nbrcaract           1.540e-08  3.529e-09   4.363 2.34e-05 ***
dureeminute:nbrchiffre      -2.284e-03  1.019e-03  -2.241 0.026489 *
dureeminute:nbrcaract       -3.151e-04  1.036e-04  -3.040 0.002780 **
nbrcaractspeciaux:nbrchiffre -3.071e-02  2.235e-02  -1.374 0.171463
nbrchiffre:nbrcaract         5.727e-03  3.487e-03   1.642 0.102576
nbrmaj:nbrcaract            -4.583e-04  2.229e-04  -2.056 0.041481 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.316439)

    Null deviance: 146.134  on 167  degrees of freedom
Residual deviance:  48.415  on 153  degrees of freedom
AIC: 299.75

Number of Fisher Scoring iterations: 2
```

Conclusion :

Afin de faire un maximum de vue sur un podcast il faut préviligier des titres long comprenant des majuscules et augmenter la durée de la vidéo