TikTok

VS

- Bloc Transversal -

Resum

L'objectiu principal d'aquest estudi és comparar els algoritmes de recomanació als nous usuaris de les plataformes YouTube i TikTok.

Per agafar dades de manera aleatòria ens hem registrat a aquestes amb un nou correu electrònic, i durant diferents dies hem recollit dades sobre els vídeos que se'ns recomanaven.

Hem creat un model lineal per a poder fer prediccions a partir d'aquestes dades i hem pogut comprovar que un vídeo amb les mateixes característiques a ambdues plataformes té més visualitzacions a TikTok que a YouTube.

L'aleatorietat de l'estudi no es pot assegurar al 100% perquè hi ha tota una sèrie de factors que alteren a la recollida de dades, com per exemple la ubicació dels nostres dispositius o els esdeveniments importants que coincideixen amb els dies que s'agafen. Per aquesta mateixa raó, és un estudi poc reproduïble i generalitzable.

Introducció

En un món cada cop més connectat, el consum de contingut audiovisual s'ha convertit en un fenomen fonamental en l'experiència digital. Plataformes com YouTube i TikTok han emergit com a gegants en la difusió de vídeos, modelant tendències i redefinint com les audiències interactuen amb el contingut. Davant la prevalença d'aquestes plataformes, sorgeix la inquietud de comprendre com seleccionen i recomanen vídeos als seus nous usuaris.

La motivació fonamental d'aquest estudi és indagar en els algoritmes de recomanació de YouTube i TikTok, enfocant-nos en un aspecte específic: quin tipus de vídeos són prioritzats per als nous usuaris? Aquesta recerca sorgeix de la necessitat de comprendre el criteri que aquestes plataformes empren en presentar contingut als seus usuaris no familiaritzats, i d'aquesta forma intentar apropar-nos al comportament real d'aquestes plataformes que la gran majoria de la població utilitza o ha utilitzat algun cop a la seva vida i de les quals no tenim massa informació respecte al contingut que se'ns recomana.

El present estudi té com a objectiu principal analitzar i comparar el comportament de les plataformes YouTube i TikTok en relació amb la promoció de vídeos als nous usuaris, en específic, analitzar quina d'aquestes dues grans plataformes recomana vídeos amb una ràtio major de visualitzacions en funció del temps antiguitat del vídeo a la plataforma. Són realment aleatoris els algoritmes de recomanació de YouTube i TikTok? Quina prioritza vídeos més populars i quina és més imparcial? Mitjançant aquest estudi arribarem a una resposta a aquestes preguntes.

Partint de l'experiència prèvia que tenim respecte a aquestes dues plataformes arribem a una hipòtesi inicial:

TikTok inclina el seu algoritme de recomanació cap a vídeos amb una ràtio de visualitzacions/temps del vídeo a la plataforma superior al de YouTube i, per tant, un vídeo a TikTok arribarà a les mateixes visualitzacions que a YouTube en un temps menor.

Mètodes

Variables

La nostra variable **X** seran les dues plataformes que volem comparar (YouTube o Tik Tok).
La nostra variable d'interès **Y** serà la ràtio de visualitzacions/temps d'antiguitat a la plataforma del vídeo, que mesurarem en visualitzacions/minut. A més, recollirem dues variables **Z** que seran likes/temps d'antiguitat a la plataforma del vídeo, mesurada en likes/minut, i nombre de subscriptors (en el cas de YouTube) o seguidors (en el cas de TikTok) que té el creador del vídeo.

Recollida de dades

Respecte a la recollida de dades el procediment que seguirem serà el següent:
Crearem un compte nou de correu electrònic (en el nostre cas crearem un compte de Gmail de Google) i ens registrarem a les plataformes YouTube i TikTok. Entrarem cada dia durant 6 dies seguits a ambdues plataformes, 2 dies cada integrant del grup i, per tant, la recollida de dades es realitzarà des de 3 dispositius i 3 xarxes diferents, i recollirem les dades de 5 vídeos de cada plataforma, els que ens recomanin. D'aquesta manera recollirem les dades de 60 vídeos en total, 30 d'una plataforma i 30 de l'altre, en un interval de 6 dies.

Per garantir la màxima aleatorietat dels resultats seguirem les següents mesures:
Per començar el compte que utilitzarem tant a YouTube com a TikTok serà un compte nou, amb un correu electrònic nou (en el nostre cas serà de Google).
Les dades seran recollides durant un interval prolongat de temps evitant així alteracions en els algoritmes de recomanació de vídeos deguts a successos puntuals de gran repercussió com vindria a ser un partit de futbol. A més, per garantir l'aleatorietat dels vídeos recomanats emprarem tres dispositius i tres xarxes diferents per a l'estudi i sempre recollint les dades sense interactuar amb els vídeos recomanats per evitar així possibles cookies que alterarien l'algoritme de recomanació.
D'aquesta manera aconseguirem una mostra aleatòria en la màxima mesura del que podem, i tot i que realitzarem la recollida de dades des de tres localitzacions diferents, en ser totes tres en Catalunya no podrem evitar el factor de la localització, que potser influeix en l'aleatorietat de les dades.

Cal remarcar que les dades de les quals disposem són independents, ja que compararem vídeos diferents entre les dues plataformes.

Anàlisi estadística

Per a fer l'interval de confiança de les diferències de les mitjanes poblacionals usem les següents instruccions (com les dades de la variable resposta per si soles no compleixen la normalitat vista amb els gràfics qqnorm i qqline, traiem logaritme de la variable resposta Y i desfem el canvi al final:

- t.test(log(datos$Y) ~ datos$X, var.equal = TRUE)
- resultat_log <- c(0.5835631,3.1248972)
- resultat <- exp(resultat_log)
- resultat

Per a la construcció d'un model lineal cal veure si compleix les premisses (en el nostre cas per a un model lineal múltiple):

- Linealitat
- Normalitat
- Homoscedasticitat
- Independència

I per comprovar-les utilitzarem les següents instruccions:

- par(mfrow=c(2,2)) #per la linealitat i homoscedasticitat
- plot(mod_mult2,c(2,1)) # per la normalitat
- plot (rstandard(mod_mult2),type="l")# per la independència de les dades

Podem avançar que les premises del model normal no és compleixen, i per tant farà falta construir el model amb logaritmes en les dades i tornarem a comprovar totes les premises.

El model que utilitzem és el múltiple del tipus:
$\log(Y)=\beta 0+\beta 1\times as.factor(X)Youtube+\beta 2\times \log(Z1)+\beta 3\times \log(Z2)+\beta 4\times \log(Z3)$ +error

I les instruccions en R:

- mod_mult2 <- lm(log(Y) ~ as.factor(X) + log(Z1) + log(Z2) + log(Z3),datos)
- summary(mod_mult2)

Resultats

Descriptiva de les dades

Les següents dades i gràfics mostren la descriptiva del paràmetre resposta Y (visualitzacions/temps en la plataforma) segons el grup del paràmetre X al qual pertanyen els vídeos (TikTok o Youtube):

Visualitzacions/Temps en la plataforma en Tiktok

Min	Max	Mitjana	Mediana	1r Quantil[1]	3r Quantil[2]
9.428	2640.999	433.031	293.488	36.197	573.842

Variància: 305275.1
Desviació tipus: 552.5170

Visualitzacions/Temps en la plataforma en Youtube

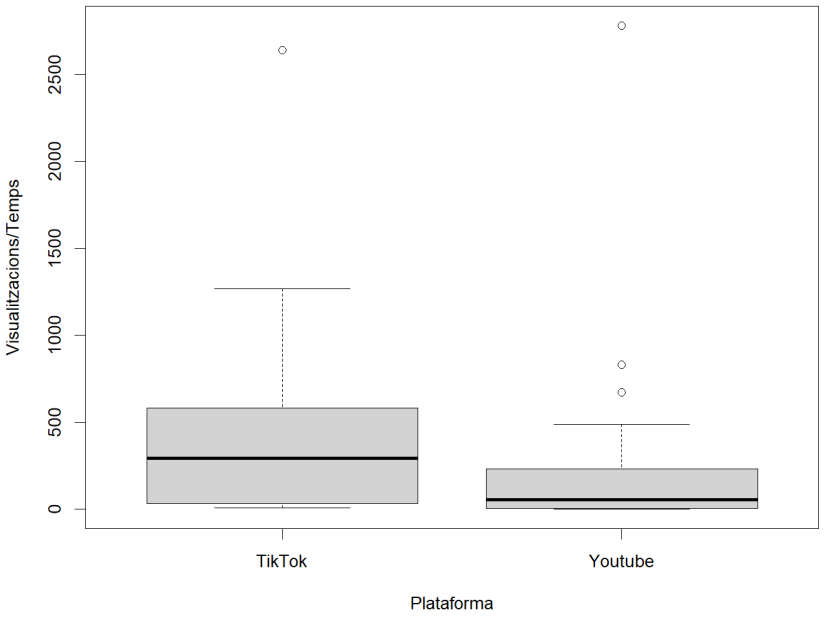
Min	Max	Mitjana	Mediana	1r Quantil	3r Quantil
0.011	2781.428	245.830	54.372	3.677	234.075

Variància: 286173.8
Desviació tipus: 534.9521

Amb aquestes dades podem veure com els 30 videos de Tiktok respecte dels 29 videos de Youtube (hem hagut de treure'n 1 ja que era molt atípic, explicat a l'annex) tenen una mitjana de visualitzacions/temps en la plataforma molt més gran (gairebé el doble), cosa que també es pot veure amb la mediana, on podem veure com en els 2 casos és molt diferent de la mitjana, cosa que pot indicar valors molt extrems.
La variabilitat i desviació estàndard és semblant en els dos casos.

[1] Valor el qual el 25% de les dades de Y es troba per sota en aquella categoria
[2] Valor el qual el 75% de les dades de Y es troba per sota en aquella categoria

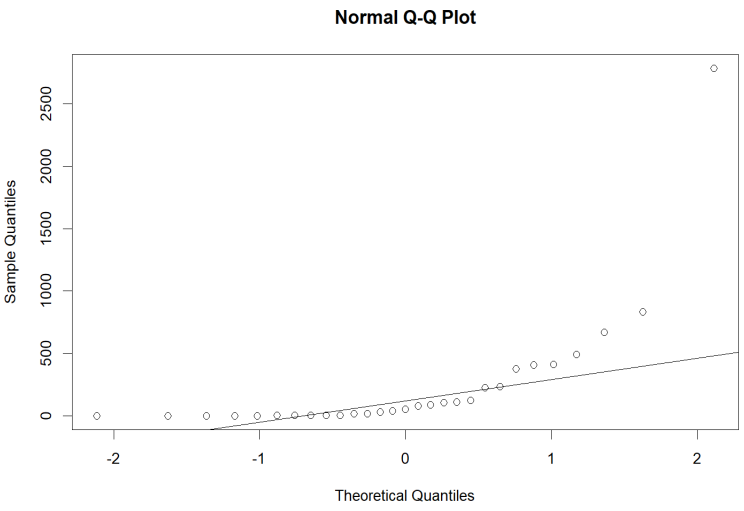
Boxplot Visualitzacions/Temps vs Plataforma

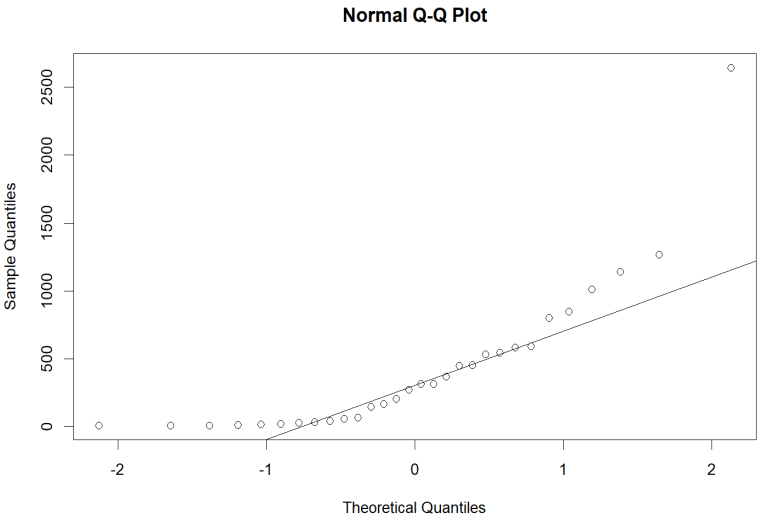


Podem veure el que ja havíem comentat, la mediana de les visualitzacions/temps a la plataforma és superior a Tiktok que a Youtube, ja que la línea negra de la caixa de Tiktok es troba més a dalt. A més podem veure con Youtube té més valors atípics, ja que tenim més punts fora dels bigotis.

Premisa: Normalitat en les dades

Comprovem amb el gràfic qqnorm() i qqline() si les dades Y en funció de X són normals o no i si cal fer transformacions logarítmiques.

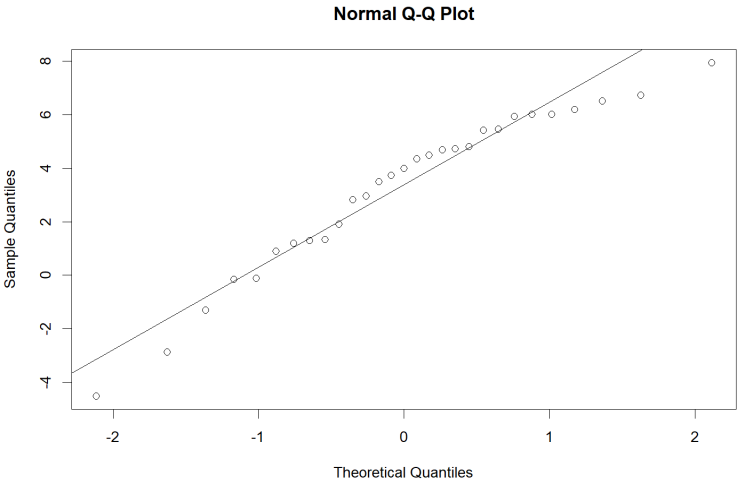




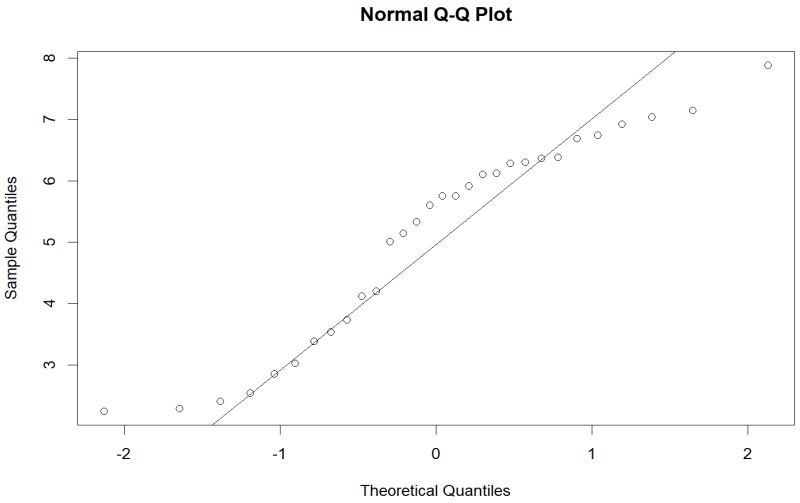
Youtube

Tiktok

Veient la NO normalitat, treiem logaritmes de les dades per a tenir dades normals:



Youtube (logaritme)



Tiktok(logaritme)

Veiem en els gràfics que amb logaritmes en la variable Y obtenim dades normals (no exactament, però el gràfic ha millorat molt respecte de l'anterior)

Càlcul interval de confiança de diferència de mitjanes

Amb la instrucció t.test, calcularem un interval de confiança de la diferència de mitjanes entre dels grups del 95%. Com hem avançat en l'apartat anterior usarem dades logarítmiques, per tant, caldrà desfer el canvi un cop calculat l'IC.

IC(log(μ1/ μ2), 95%) = [0.5835631 3.1248972]
IC(μ1/μ2, 95%) = [e^0.5835631 e^3.1248972] = [1.792414,22.757555]

D'aquest resultat podem treure que el quocient entre les mitjanes de les poblacions de Tiktok i Youtube no serà inferior a 1.792414 amb un 95% de confiança, tampoc serà superior a 22.757555 amb un 95% de confiança, i veiem que com el valor 1 no és plausible, sabem que amb un 95% de confiança que la mitjana poblacional visualitzacions/temps en la plataforma de les dues plataformes sigui igual.

Càlcul d'un model lineal múltiple per a fer prediccions

Trobem el model lineal múltiple que engloba totes les variables de les dades recollides. Amb l'script de R summary(lm(Y ~ as.factor(X) + Z1 + Z2 + Z3,datos)), podem saber els paràmetres b0,b1,b2,b3 que defineixen el model on ens queda el següent model amb l'output de la imatge (output reduït):

Model 1 : $Y=412.1-193.2\times Youtube-0.00003801\times Z1+0.00001292\times Z2-0.1653\times Z3$

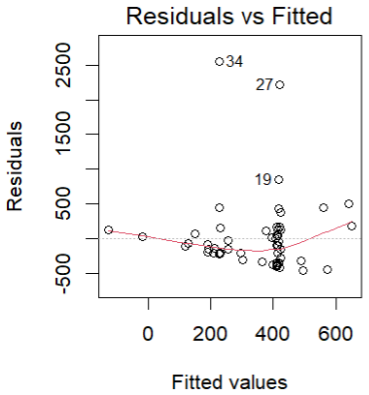
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.121e+02  1.015e+02   4.059  0.00016 ***
as.factor(X)Youtube -1.932e+02  1.594e+02  -1.212  0.23092
Z1                  -3.801e-05  2.001e-04  -0.190  0.85002
Z2                   1.292e-05  1.099e-05   1.176  0.24480
Z3                  -1.653e-01  2.282e-01  -0.724  0.47191
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 547.1 on 54 degrees of freedom
Multiple R-squared:  0.07001,   Adjusted R-squared:  0.001126
F-statistic: 1.016 on 4 and 54 DF,  p-value: 0.4072
```

Veiem que el model només és capaç d'explicar un 7% de la variabilitat. Igualment fem la comprovació de les premisses.

Comprovem les premises

Linealitat i Homoscedasticitat:

Veiem que aquest model NO compleix les premisses de linealitat i homoscedasticitat, apliquem transformacions logarítmiques a totes les variables.



Model amb logaritmes

Usem el mateix script que abans però traient logaritmes a les dades. La comanda és summary(lm(log(Y) ~ as.factor(X) + log(Z1) + log(Z2) + log(Z3),datos)) on ens queda el següent model.

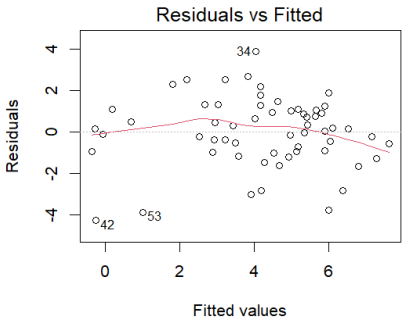
Model2 :$\log(Y)=-0.71503-1.77890\times as.factor(X)Youtube+0.17343\times \log(Z1)$ $+0.44078\times \log(Z2)-0.07359\times \log(Z3) + 1.727$

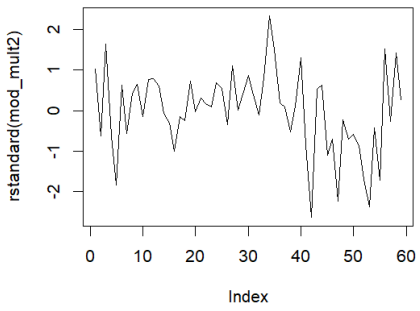
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.71503    0.90333  -0.792   0.4321
as.factor(X)Youtube -1.77890    0.96741  -1.839   0.0714 .
log(Z1)              0.17343    0.06899   2.514   0.0150 *
log(Z2)              0.44078    0.07602   5.799 3.58e-07 ***
log(Z3)             -0.07359    0.17957  -0.410   0.6835
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.727 on 54 degrees of freedom
Multiple R-squared:  0.5864,    Adjusted R-squared:  0.5557
F-statistic: 19.14 on 4 and 54 DF,  p-value: 7.501e-10
```

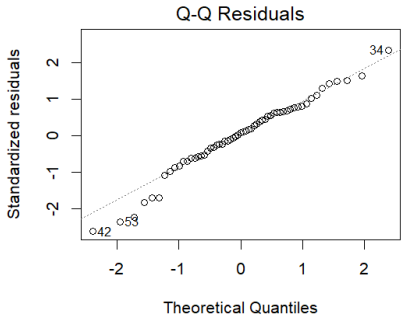
On veiem que el model explica un 58% de la variabilitat, per tant, ha millorat força respecte al model sense logaritmes.

Comprovem les premises amb el model amb logaritmes

Linealitat i Homoscedasticitat **Independència de les dades** **Normalitat**







Veiem com el model compleix més o menys la linealitat i l'homoscedasticitat, compleix també parcialment la normalitat i compleix bé la independència de les dades.

Interpretació del model

Intercept és el valor de la resposta si la resta fossin 0, en aquest cas seria les visualitzacions/temps en la plataforma de TikTok (recordem que hem tret logaritmes en el model). En el nostre cas no té sentit que Z3 fos 0, ja que no podem tenir un vídeo amb 0 minuts de duració.

Interpretem els coeficients del model.

- El coeficient de Youtube és l'únic que no porta log en ser el coeficient d'una variable categòrica, en aquest cas el coeficient de -1.77890 indica que el fet de que el video sigui de Youtube, fa disminuir el log(Y) en -1.77890.
  Quan desfem la transformació tenim que el fet de ser de Youtube fa disminuir la Y en exp(−1.77890), que és 0.169. Per tant, les visualitzacions/temps dels vídeos de Youtube son un 83% inferiors a les de Tiktok.

Els coeficients de les variables contínues si tenen logaritmes, els tractem. Veurem que passa si multiplico la variable Z1 * 2, volem veure el canvi percentual en la variable Y.

Tenim:

- log(Y1)=−0.71503−1.77890×as.factor(X)Youtube+0.17343×log(Z1)+0.44078×log(Z2) − 0.07359×log(Z3)
- log(Y2)=−0.71503−1.77890×as.factor(X)Youtube+0.17343×log(2*Z1)+0.44078×log(Z2) − 0.07359×log(Z3)

Si restem les equacions ens queda, per les propietats dels logaritmes:

- log(Y2) - log (Y1) = 0.17343 * [log(2*Z1) - log(Z1)] -->
  Implica que log(Y2/Y1) = log(2^0.17343) --> Y2/Y1 = 2^0.17343 = 1.128

Per tant, si multipliquem * 2 Z1 obtenim que el nombre de visualitzacions/temps(variable Y) s'incrementa en un 12.8%. Podem repetir el càlcul per veure quan s'incrementa percentualment i quan multipliquem per 3,4,5…

Repetim els càlculs per a Z2

- log(Y1)=−0.71503−1.77890×as.factor(X)Youtube+0.17343×log(Z1)+0.44078×log(Z2) − 0.07359×log(Z3)
- log(Y2)=−0.71503−1.77890×as.factor(X)Youtube+0.17343×log(Z1)+0.44078×log(2*Z2) − 0.07359×log(Z3)

Si restem les equacions ens queda, per les propietats dels logaritmes:

- log(Y2) - log (Y1) = 0.44078 * [log(2*Z2) - log(Z2)] -->
  Implica que log(Y2/Y1) = log(2^0.44078) --> Y2/Y1 = 2^0.44078 = 1.357

Per tant, si multipliquem * 2 Z2 obtenim que el nombre de visualitzacions/temps (variable Y) s'incrementa en un 13.6%.

Finalment ho veiem pel coeficient de Z3 que està restant

- log(Y1)=−0.71503−1.77890×as.factor(X)Youtube+0.17343×log(Z1)+0.44078×log(Z2) − 0.07359×log(Z3)
- log(Y2)=−0.71503−1.77890×as.factor(X)Youtube+0.17343×log(Z1)+0.44078×log(Z2) − 0.07359×log(2*Z3)

Si restem les equacions (la primera menys la segona) ens queda, per les propietats dels logaritmes:

- log(Y1) - log (Y2) = 0.07359 * [log(2*Z3) - log(Z3)] -->
  Implica que log(Y1/Y2) = log(2^0.07359) --> Y1/Y2 = 2^0.07359 = 1.052

Per tant si multipliquem *2 Z3 obtenim que el nombre de visualitzacions/temps es disminueix en un 10.52%.

Prediccions puntuals

Farem una predicció individual i una de valor esperat
En la predicció individual predirem el nombre de visualitzacions/ temps en la plataforma que tindrà un creador de contingut de Youtube i a Tiktok amb 10000 likes/temps, 5000 seguidors i la duració del vídeo és de 5 minuts en ambdues plataformes

- predict(mod_mult2,data.frame(X = 'Youtube', Z1 = 10000, Z2 = 5000 ,Z3 =5,interval = 'prediction')
- predict(mod_mult2,data.frame(X = 'Tiktok', Z1 = 10000, Z2 = 5000 ,Z3 =5),interval = 'prediction')

En la primera predicció individual tenim 15.47418 (exp(2.739173) visualitzacions/temps. També tenim un interval de confiança del 95% on un vídeo de Youtube amb 10.000 likes/temps, 5000 subscriptors i 5 minuts de duració tindrà entre [0.3520384,680.1817215] visualitzacions/temps.
En la segona predicció individual tenim 91.65908 exp(4.518076) visualitzacions/temps. Tenim un interval de confiança del 95% que un vídeo de TikTok amb 10.000 likes/temps, 5000 subscriptors i 5 minuts de duració tindrà entre [1.928689,4356.010959] visualitzacions/temps.

Fem una predicció de valor esperat per als vídeos de YouTube amb les mateixes característiques anteriors:

- predict(mod_mult2,data.frame(X = 'Youtube', Z1 = 10000, Z2 = 5000 ,Z3 =5,interval = ‘confidence’)

La predicció puntual és la mateixa que en la predicció de valor individual, que és 15.47418, però en l’interval de confiança ha canviat, ara dona després de treure el logaritme [3.361559,71.231918].

Discussió

Interpretació dels resultats de les prediccions

Nosaltres construïm el model lineal múltiple per a poder fer prediccions de valors. En el nostre cas si volguessin saber el nombre de visites que tindrà el meu video a Youtube, si tinc 10000 likes/ temps que porti en la plataforma, tinc 5000 seguidors i el meu video dura 5 minuts, usarem la predicció individual, que hem donat que el meu video tindra entre [0.3520384,680.1817215] visualitzacions/temps amb un 95% de confiança i dona una predicció puntual de 15.47418 = (exp(2.739173)).
En canvi si vull saber el nombre de visites que tindran tots els videos que tinguin 10000 likes/temps en la plataforma, que el seu creador tingui 5000 seguidors i el seu video duri 5 minuts usarem la predicció de valor esperat, que ens dona que els videos tindran entre [3.361559,71.231918] visualitzacions/temps en la plataforma i dona una predicció puntual del valor esperat de 15.47418 (igual que en la predicció de valor individual).

Partint d'aquestes prediccions, podem valorar la nostra hipòtesi inicial. El màxim de visualitzacions/temps és molt més gran a TikTok que a YouTube i la predicció puntual també és major a TikTok, aleshores podem comprovar que la nostra primera suposició, on prediem que TikTok recomana vídeos amb una ràtio de visualitzacions/temps més elevat, era certa.

Limitacions de l'estudi

La limitació principal de l'estudi ha sigut la falta d'aleatorietat a l'hora de fer la recollida de dades. Com que els algoritmes de recomanació no són aleatoris, i es basen en factors com l'edat, el gènere i la ubicació de l'usuari, i canvien en funció de l'hora i del dia de la setmana, l'època i els esdeveniments importants tant nacionals com internacionals, és molt complicat fer-ne una recollida completament aleatòria, tot i que s'ha intentat evitar al màxim aquests elements, mitjançant una sèrie de mesures de seguretat ja exposades prèviament, poden haver acabat afectant les recomanacions que apareixien, encara que això no ho podem saber.

Generabilitat

Com que no hem pogut evitar completament els factors que afecten l'aleatorietat de l'estudi (localització geogràfica), només és extrapolable a un conjunt de població que tingui les mateixes limitacions que nosaltres i que repetís l'estudi en les mateixes condicions, intentant evitar al màxim les discrepàncies amb l'aleatorietat.

Annex 2 Script R

```
#R script Bloc transversal
#Primer eliminem tots els objectes guardats en la memòria del R  (opcional)

 rm(list = ls())

#Carreguem les dades al R i li assignem a la variable "datos" les dades, dades carregades desde
#file, import DataSet, From Excel (ja que les nostres dades són d'aquest format) i busquem
#on hem guardat les dades

 datos<- Dades_PE_Net_

#En el nostre cas, avancem que tenim un vídeo amb dades atípiques que perjudiquen els gràfics,
#així que per a tenir gràfics coherents eliminem el video.

 datos <- datos[-which.max(datos$Y),]

#El nostre treball té dades independents ja que comparem videos diferents en les diferents plataformes.
#Era impossible agafar dades aparellades ja que necessitaríem comparar els mateixos videos en les diferents plataformes,
#cosa que és gairebé impossible.

#Fem la descriptiva numérica (variables X i Y)

#-------------------------------------------------
#Mitjana de les visualitzacions/temps en la plataforma de Tik Tok i Youtube
 tapply(datos$Y,datos$X,mean)

#Mediana de les visualitzacions/temps en la plataforma de Tik Tok i Youtube
 tapply(datos$Y,datos$X,median)

#Variabilitat de les visualitzacions/temps en la plataforma de TikTok i Youtube
 tapply(datos$Y,datos$X,var)
```

```
#Standard deviation (desviació estàndard, quant s'allunyen les dades de la mitjana)de les visualitzacions/temps en la plataforma de Tik Tok i Youtube
 tapply(datos$Y,datos$X,sd)

#Summary (que dona el valor mínim,el màxim,el 1r i 3r quantil, la mediana i mitjana)
 tapply(datos$Y,datos$X,summary)

#Fem la descriptiva gràfica les dades (variables X i Y)
#-----------------------------------------------------
#Distribució de la variable Y funció de Youtube i de Tiktok (opcionalment li canviem el noms als eixos)

 boxplot(Y~X,datos)
 boxplot(Y ~ X, data = datos, xlab = "Plataforma", ylab = "Visualitzacions/Temps")


#Amb el resultat veiem que en els videos de Youtube tenim més valors atípics (3) que en Tiktok que en tenim 1
#Veiem que la variabilitat de les dades és major en Tiktok, cosa que hem pogut comprovar abans amb la descriptiva
#numérica.

#Veiem si es compleix o no la normalitat de les dades (com són dades independents primer mirarem la normalitat d'un grup i després de l'altre)
#-----------------------------------------------------------------------------------------------
-----------------------------------------------

 qqnorm(datos$Y[datos$X=="Youtube"])
 qqline(datos$Y[datos$X=="Youtube"])

 qqnorm(datos$Y[datos$X=="TikTok"])
 qqline(datos$Y[datos$X=="TikTok"])

#Veiem que NO compleixen la normalitat, treiem logaritmes per veure si millora
 qqnorm(log(datos$Y)[datos$X=="Youtube"])
 qqline(log(datos$Y)[datos$X=="Youtube"])

 qqnorm(log(datos$Y)[datos$X=="TikTok"])
 qqline(log(datos$Y)[datos$X=="TikTok"])
```

#Veiem una millora ens els gràfics, encara que segueixen sense estar sobre la recta perfectament que dibuixa el qqline
#Traiem logaritmes per a fer els intervals de confiança de les mitjanes.

#Calculem un interval de confiança del 95% per a les diferencies de les mitjanes dels 2 grups
#----------

#Amb aquesta instrucció estem calculant un interval de confiança de la resta de mitjanes μ1- μ2.
#Com traiem logarimtes de les dades, quan desfem el canvi haurem d'elevar [e^extrem_inferior_interval,e^extrem_superior_interval]

```
t.test(log(datos$Y) ~ datos$X, var.equal = TRUE)
resultat_log <- c(0.5835631,3.1248972 )
resultat <- exp(resultat_log)
resultat
```

#El resultat ha sortit positiu i el valor 1 no es un valor plausible en l'interval de confiança
#cosa que indica que NO es plausible que les dues mitjanes siguin iguals.
#Com Tiktok representa μ1 i Youtube representa μ2, podem afirmar amb un 95% de seguretat que la mitjana
de visualitzacions/temps en la plataforma és més gran entre [1.792414, 22.757555] que Youtube.

#Ajustem un model lineal múltiple amb les nostres dades per intentar fer prediccions amb el model
#----------
--

#Amb el nostre model intentarem predir el nombre de visualitzacions/temps en la plataforma que tindrà un video
#el nombre de seguidors en la plataforma, la seva duració y la seva quantitat de likes/visualitzacions

#Model lineal múltiple amb totes les variables

```
mod_mult1 <- lm(Y ~ as.factor(X) + Z1 + Z2 + Z3,datos)
summary(mod_mult1)
```

```
#Veiem si el model compleix les premises dels models lineals simples/múltiples
#Les premises que hem de veure són la linealitat, la normalitat, la homocedasticitat i la independència de les dades

 par(mfrow=c(2,2))
 plot(mod_mult1,c(2,1)) # QQ-Norm i Standard Residuals vs. Fitted
 hist(rstandard(mod_mult1),font.main=1) # Histograma dels residus estandaritzats
 plot (rstandard(mod_mult1),type="l")

#Després de fer crear el model veiem que el R^2 es de un 7%, cosa que indica que el
#model explica fatal la variable resposta, i també veiem com no compleix les premises de
#homoscedasticitat i normalitat. Fem transformacions logarítmiques per veure si millora el model

 mod_mult2 <- lm(log(Y) ~ as.factor(X) + log(Z1) + log(Z2) + log(Z3),datos)
 summary(mod_mult2)



#Comprovem les premises
 par(mfrow=c(2,2))
 plot(mod_mult2,c(2,1)) # QQ-Norm i Standard Residuals vs. Fitted
 hist(rstandard(mod_mult2),font.main=1) # Histograma dels residus estandaritzats
 plot (rstandard(mod_mult2),type="l")

#Veiem que el nou model té un R^2 del 58%, cosa que és bastant millor encara que no perfecte
#De les premises,les dades no compleixen del tot la normalitat, cosa que ja hem vist en l'apartat
#de la descriptiva numérica.
#Com a variables explicatives agafem els likes/temps en la plataforma (Z1), nombre de subscriptors (Z2)
# i nombre de subscriptors/seguidors (Z3), i també la variable si es de Youtube o no (X)
#Tiktok esta inclós en el intercept inicial. Amb aquestes variables predictores del model
#s'intenta predir el nombre de visualitzacions/temps en la plataforma

#Intepretació de les dades del model

#Prediccions de valor Individual o de valor esperat y que signifiquen
```

```
  predict(mod_mult2,data.frame(X = 'Youtube', Z1 = 10000, Z2 = 5000 ,Z3 =5 ),interval =
'prediction')
   predict(mod_mult2,data.frame(X = 'TikTok', Z1 = 10000, Z2 = 5000 ,Z3 =5),interval =
'prediction')
  predict(mod_mult2,data.frame(X = 'Youtube', Z1 = 10000, Z2 = 5000 ,Z3 =5 ),interval =
'confidence')
```