TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 2

Ferran Gurri Mancera

17/05/2018

Índex

1. Descripció del dataset

Per l'elaboració d'aquesta pràctica he triat un dels datasets públics disponibles al portal https://www.kaggle.com suggerit en l'enunciat, concretament el que fa referencia als videos de Youtube que són trending tòpic a Canadà (https://www.kaggle.com/datasnaek/youtube-new).

He descartat el dataset elaborat en la pràctica 1 perquè he considerat que no tenia prou dades numèriques per poder realitzar-ne una neteja de dades complerta ni tampoc un estudi estadístic significatiu.

El dataset triat consta dels 200 videos trending tòpic de cada país enregistrat dia a dia i ens permetrar fer un estudi de quines són les categories que més impacte tenen en la xarxa social basada en videos més extesa de internet.

L'objectiu d'aquest projecte és analitzar els videos que arriben a ser trending tòpic (de més interés per la comunicat basat en el nombre de visualitzacions, comentaris, likes, etc) en funció de la seva categoria. Busquem doncs donar resposta a preguntes com ¿És Youtube una plataforma principalment de música? ¿En quina categoria ens hauríem d'especialitzar si volguessim dedicar-nos professionalment a la pujada de videos? I, adicionalment, ¿Existeix una relació entre el nombre de visites i el número de likes, dislikes i número de comentaris?

2. Integració i selecció de les dades d'interès a analitzar

El dataset ve amb un conjunt considerable de columnes incloses i no totes ens aporten informació útil al problema que volem analitzar. Per fer una primera ullada a les dades descarregades les obrirem en format Excel i n'analitzarem el contingut. Complementarem aquesta informació amb les metadates de les columnes que disposem en el portal que conté el dataset original.

Esquematitzaré aquesta informació en la següent taula a mode de guía inicial en el procés de neteja:

Camp	Format	Descripció	Ens interessa?
video_id	alfanumeric	Codi identificador del video	Sí
trending_date	Data: YY.DD.MM	Dia en que va ser trending topic el video.	Sí
title	alfanumeric	Títol del video publicat	No
channel_title	Alfanumeric	Nom del canal que conté el video	No
category_id	Numeric	Codi identificador de la categoria	Sí
publish_time	Alfanumeric	Dia i hora en que es va publicar el video	No
Tags	Alfanumeric	Llista de tags definits per l'usuari.	No

		Venen separats per "|"	
Views	Numeric	Nombre de visites que ha rebut el video	Sí
likes	Numeric	Nombre de "m'agrada" que ha rebut el video	Sí
dislikes	Numeric	Nombre de "No m'agrada" que ha rebut el video	Sí
comment_count	Numeric	Nombre de comentaris que conté el video	Sí
thumbnail_link	Alfanumeric	Enllaç a la vista prèvia del video	No
comments_disabled	Alfanumeric	"False" o "True" indicant si el video permet comentaris	Sí en cas de que volguem fer análisis sobre el nombre de comentaris del video
ratings_disabled	Alfanumeric	"False" o "True" indicant si el video permet valoracions	Sí, en cas de que volguem fer anàlisi sobre el nombre de likes/dislikes
video_error_or_removed	Alfanumeric	"False" o "True" indicant si el video ha estat eliminat (en aquest cas es conserven les dades que tenia en el moment de ser eliminat)	Sí, ja que pot esbiaixar les dades
description	Alfanumeric	Text obert on la persona que ha pujat el video n'escriu una descripció	No

Per tant el nostre dataset constarà de les columnes: video_id, trending_date, category_id, views, likes, dislikes, comment_count, comments_disabled, ratings_disabled , video_error_or_removed

Algunes consideracions prèvies a destacar son:

- "trending_date" pren valors entre el 16 de novembre del 2017 (17.16.11) i el 7 d'abril del 2018
- el camp "category_id" és un identificador numèric intern, ens donen la seva equivalència en un fitxer CA_category_id.json a part.
- El camp de text obert "description" conté caracters que trenquen l'estructura en linees del fitxer i caldrà revisar que es carreguen correctament
- Es probable que existeixi relació entre les columnes que enregistren el nombre de visites amb el numbre de "likes", "dislikes" o "comment_count" ja que cal pensar que a més visites més probabilitat de tenir-ne activitat.
- Els camps "comments_disabled", "ratings_disabled" i "video_error_or_removed" ens poden aportar valors zero o incomplerts i requerirant un tractament a part si no volem que ens esbiaixin les dades. És a dir, si per exemple volem analitzar la distribució per nombre de comentaris realitzats i no tenim en compte si en el video els comentaris estan permesos estarem incloent moltes dades a 0 incorrectament. Un cas més

problemàtic és el camp “video_error_or_removed”, ja que si ens indica que el video va ser retirat llavors totes les dades fan referencia al temps que va estar pujat i no sabrem quines serien les seves dades totals en cas de no haver estat retirat.

- Per simplicitat de l’estudi, no tindrem en compte el camp “publish_time”, però podria ser interessant intentar relacionar els dies que porta publicat el video amb l’activitat rebuda
- Tampoc analitzarem el camp “tags” perquè volem fer l’analisi en funció de les categories. Una possible ampliació del estudi de les dades seria tallar la llista de tags rebudes en aquest camp i repetir-ne l’analisi prenent-les com una categorització alternativa.
- Amb unes transformacions bàsiques en Excel comprovem el nombre de videos per dia inclosos en les dades. En la descripció del dataset se’ns informa que s’han seleccionat els 200 videos más trending topic per dia, però observem que hi ha dies incomplerts:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1																		
2	18.01.02	190	17.16.12	199	18.22.02	199	17.15.11	200	17.26.12	200	18.03.05	200	18.09.01	200	18.18.01	200	18.25.01	200
3	18.31.01	191	17.19.11	199	18.23.04	199	17.15.12	200	17.27.11	200	18.04.01	200	18.09.02	200	18.18.03	200	18.25.02	200
4	18.02.02	195	17.21.12	199	18.26.02	199	17.16.11	200	17.27.12	200	18.04.02	200	18.10.03	200	18.18.04	200	18.25.03	200
5	18.13.02	195	17.26.11	199	18.27.01	199	17.17.11	200	17.28.11	200	18.04.04	200	18.11.03	200	18.19.02	200	18.26.01	200
6	18.29.01	196	17.29.12	199	18.27.03	199	17.17.12	200	17.28.12	200	18.04.05	200	18.12.01	200	18.19.04	200	18.26.03	200
7	18.09.03	197	18.03.02	199	18.28.03	199	17.18.11	200	17.29.11	200	18.05.01	200	18.12.02	200	18.20.01	200	18.26.04	200
8	18.19.03	197	18.03.03	199	18.29.04	199	17.18.12	200	17.30.11	200	18.05.03	200	18.12.03	200	18.20.02	200	18.27.04	200
9	18.25.04	197	18.04.03	199	17.01.12	200	17.19.12	200	17.30.12	200	18.05.04	200	18.13.01	200	18.20.03	200	18.28.01	200
10	18.28.02	197	18.05.02	199	17.02.12	200	17.20.11	200	17.31.12	200	18.05.05	200	18.13.03	200	18.20.04	200	18.28.04	200
11	17.07.12	198	18.07.03	199	17.03.12	200	17.20.12	200	18.01.01	200	18.06.01	200	18.14.01	200	18.21.01	200	18.29.03	200
12	18.06.02	198	18.14.02	199	17.05.12	200	17.21.11	200	18.01.03	200	18.06.03	200	18.15.01	200	18.21.03	200	18.30.01	200
13	18.10.02	198	18.14.03	199	17.06.12	200	17.22.11	200	18.01.04	200	18.06.04	200	18.15.03	200	18.21.04	200	18.30.03	200
14	18.11.02	198	18.14.04	199	17.08.12	200	17.22.12	200	18.01.05	200	18.06.05	200	18.15.04	200	18.22.03	200	18.30.04	200
15	18.15.02	198	18.16.02	199	17.09.12	200	17.23.11	200	18.02.01	200	18.07.01	200	18.16.01	200	18.22.04	200	18.31.03	200
16	18.23.02	198	18.17.02	199	17.10.12	200	17.23.12	200	18.02.03	200	18.07.02	200	18.16.03	200	18.23.01	200		
17	18.24.02	198	18.18.02	199	17.11.12	200	17.24.11	200	18.02.04	200	18.07.04	200	18.16.04	200	18.23.03	200		
18	18.27.02	198	18.19.01	199	17.13.12	200	17.24.12	200	18.02.05	200	18.08.01	200	18.17.01	200	18.24.01	200		
19	17.04.12	199	18.21.02	199	17.14.11	200	17.25.11	200	18.03.01	200	18.08.02	200	18.17.03	200	18.24.03	200		
20	17.12.12	199	18.22.01	199	17.14.12	200	17.25.12	200	18.03.04	200	18.08.03	200	18.17.04	200	18.24.04	200		

Crec que és poc probable que aquests dies incomplerts es tracti de dies que no s’han arribat a publicar 200 videos en un dia. Estem per tant devant de un cas de dades buides perquè manquen registres que sabem que existeixen però no tenim.

Tractaré aquests casos en apartats següents de la pràctica.

Carregarem doncs les nostres dades mitjançant les següents instruccions en R:

Carreguem el csv:

```
> youtubes_raw <- read.csv("C:/uoc/Semestre 1/Tipologia i cicle de vida de les dades/Practica2/CAvideos.csv");
```

Seleccionem les columnes que volem fer servir en el nostre análisis:

```
> youtubes <- youtubes_raw[, c(1,2,5,8,9,10,11,13,14,15)];
```

Filtrem les columnes sense video_id informat per asegurar-nos que la càrrega no ha tingut problemes amb els registres on el camp “description” trencava l’estructura de linees:

```
> youtubes_clean <- youtubes[is.na(youtubes$video_id) == 0,];
```

Treiem un primer resum del estat de les dades després d'aquesta primera càrrega inicial:

```
> summary(youtubes_clean)
       video_id      trending_date    category_id        views              likes            dislikes         comment_count    comments_disabled rating:
 6ZfuNTqbHE8:     8   17.01.12:  200   Min.   : 1.00   Min.   :     1000   Min.   :      0   Min.   :      0.0   Min.   :     0   False:32628       False:
 l_lblj8Cq0o:     8   17.02.12:  200   1st Qu.:20.00   1st Qu.:   137261   1st Qu.:   2058   1st Qu.:     92.0   1st Qu.:   390   True :  486       True :
 UceaB4D0jpo:     8   17.03.12:  200   Median :24.00   Median :   354587   Median :   8637   Median :    289.0   Median :  1272
 9v_rtaye2yY:     7   17.05.12:  200   Mean   :20.75   Mean   :  1113774   Mean   :  38533   Mean   :   2000.9   Mean   :  4839
 Cmg2BnAv4WQ:     7   17.06.12:  200   3rd Qu.:24.00   3rd Qu.:   937967   3rd Qu.:  28194   3rd Qu.:    933.8   3rd Qu.:  3623
 doP7xKdGOKs:     7   17.08.12:  200   Max.   :43.00   Max.   :137843120   Max.   :3014479   Max.   :1602383.0   Max.   :827755
 (Other)    :33069   (Other) :31914
 video_error_or_removed
 False:33088
 True :   26
```

Observem que:

- Tenim 33114 registres
- No hi ha cap NA en cap dels camps
- Crida l'atenció que el nombre mínim de visites sigui un número rodó 1000. No he trobat confirmació sobre si només es tenen en compte videos que tinguin com a mínim 1000 visites a l'hora de considerar-los candidats a ser trending tòpic. Només se'ns diu que la API de Youtube té en compte diferents factors a l'hora de marcar un video com a Trending Tòpic. Que existís un nombre mínim de visites podria ser una explicació al fet que alguns dies no s'arribin a 200 videos.
- Existeixen videos amb 0 likes, dislikes o comment_count. Tal i com hem indicat abans, això és degut a que es poden deshabilitar a l'hora de publicar el video.

3. Neteja de les dades

3.1 Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

En el tractament previ de les dades ja hem detectat diferents casos de zeros i elements buits. Considerarem cada cas per separat.

Dies en els quals disposem de menys de 200 videos

En la descripció del dataset se'ns diu que les dades s'han generat a partir de recollir els 200 videos més trending topic per dia al Canadá. En les dades en canvi observem que hi ha dies que no tenim 200 registres. Aquest és un cas d'elements buits on no ens falta el valor de un camp sino el registre sencer que sabem que sí existeix (costa imaginar-se que en un dia no hi hagi 200 videos amb prou activitat en tot un país).

No trobo cap aclariment en la documentació sobre aquest punt en concret. La interpretació que jo faig és que la API de Youtube fa un filtre previ dels videos que té que considerar i que els videos que falten no cumpleixen algún d'aquests mínims. Una possibilitat que ja he comentat és que és probable que hi hagi un mínim de 1000 vistes per considerar un video com a candidat a ser trending tòpic.

Devant la impossibilitat de reproduir les dades que sabem que falten i tenint en compte que l'impacte d'aquesta manca de informació en l'estudi que volem realitzar és mínim prenem la mesura de ignorar aquestes dades incompletes. Ho assumirem com una limitació del sistema i informaríem a l'origen de dades perquè ho tinguin en compte.

L'altra possibilitat seria un error puntual en la recollida de dades, però tampoc tenim informació sobre el procés que captura les dades. En el cas en que les dades de un dia en concret només s'intentessin recollir una única vegada llavors podríem pensar en un error del procés i en suggeriríem que es repetís la càrrega n-vegades per solventar errors puntuals.

Likes, dislikes i comment_count a 0

Les columnes "likes", "dislikes" i "comment_count" ens permetran relacionar l'activitat realitzada sobre el video amb el nombre de visites i, per tant, són significatives pel nostre estudi. En el cas dels videos on no es registren aquestes dades perquè està deshabilitada la opció de premer el "like", el "dislike" o la possibilitat d'introduir algún comentari tindrem aquestes columnes a 0 i ens esbiaixaran els resultats. Així doncs no podem deixar aquestes dades a zero.

Les opcions disponibles devant d'aquest cas són:

- Opció 1: Eliminar aquests registres del nostre conjunt de dades
- Opció 2: Informar-los amb el valor promig dels registres que sí tenim informats
- Opció 3: Fer una estimació del valor més probable en funció del nombre de visites. És a dir, calcular la relació entre el nombre de visites i el número de likes (per exemple) i crear el ratio likes_per_visita i el mateix per cada columna

Opció 1: no tractar aquestes dades

Les dades amb valors a zero suposen un percentatge molt baix respecte al total de registres. L'estimació més pesimista seria que cada comentari o valoració bloquejada pertany a videos diferents. És a dir, que no hi ha cap cas on en un mateix video s'hi han bloquejat els comentaris i també els "m'agrada".

Així doncs el màxim de videos afectats son: 486 (comentaris bloquejats) + 245 (valoracions bloquejades) = 731 sobre un total de 33114 registres, lo que suposa un (731/33114)*100 = 2,2% del total.

Aquesta opció no seria dolenta ja que ens permetria quedar-nos nomes amb dades netes sense perdre gran quantitat d'informació.

Opció 2: informar les dades a zero amb el valor promig

Separariem les dades no zero de les zero i en calcularíem el valor promig per després actualitzar els registres amb dades zero.

Aquesta opció és bona perquè no ens modificarà els estimadors estadístics i podrem disposar de la resta de informació de les columnes ben informades.

Opció 3: demostrar la correlació entre les variables a zero i les vistes i aplicar els ratios als registres a zero per deduir-ne el resultat

És la opció més ambiciosa en el sentit en que intentem deduir el valor que ens hagués arribat en cas de no venir a zero. No obstant en aquesta fase del projecte no tenim demostrada estadísticament la correlació entre variables. Descartarem doncs aquesta opció.

Considero que la millor opció és la 2. Aquestes són les comandes en R per actualitzar les dades a zero amb el valor promig:

Primer comprovem que el subconjunt amb "ratings_disabled" a "True" té efectivament els valors "likes" i "dislikes" a 0:

```
> youtubes_clean_no_likes <- youtubes_clean[youtubes_clean$ratings_disabled == "True",];
> summary(youtubes_clean_no_likes)
        video_id     trending_date  category_id        views              likes          dislikes      comment_count    comments_disabled ratings_disabled
 BhIEIO0vaBE:   6   17.22.12:   6   Min.   : 1.00   Min.   :    2307   Min.   :0   Min.   :0   Min.   :   0.0   False:186         False:  0
 nx1R-eHSkfM:   6   18.05.02:   5   1st Qu.:24.00   1st Qu.:   54876   1st Qu.:0   1st Qu.:0   1st Qu.:   3.0   True : 59         True :245
 z3XJdMwAvZI:   5   18.19.03:   5   Median :24.00   Median :  139664   Median :0   Median :0   Median : 202.0
 -H90GPnH1q8:   4   18.04.02:   4   Mean   :21.99   Mean   : 1529910   Mean   :0   Mean   :0   Mean   : 745.3
 gGnBX6e3RJc:   4   18.24.03:   4   3rd Qu.:25.00   3rd Qu.:  709490   3rd Qu.:0   3rd Qu.:0   3rd Qu.:1031.0
 mRgq5aoPz0w:   4   18.27.02:   4   Max.   :29.00   Max.   :51243149   Max.   :0   Max.   :0   Max.   :5953.0
 (Other)    :216   (Other) :217
 video_error_or_removed
 False:245
 True :  0
```

Calculem la mitjana del valor sense aquests registres:

```
> youtubes_clean_likes <- youtubes_clean[youtubes_clean$ratings_disabled == "False",];

> like_mean = mean(youtubes_clean_likes$likes);

> like_mean
[1] 38820.15
```

Actualitzem els valors:

```
> youtubes_clean <- within(youtubes_clean, likes[ratings_disabled == "True"] <- like_mean);
```

Comprovem la mitjana després d'haver actualitzat:

```
> mean(youtubes_clean$likes);
[1] 38820.15
```

Repetim el procés però amb el camp "dislike":

```
> dislike_mean = mean(youtubes_clean_likes$dislikes);

> dislike_mean
[1] 2015.779

> youtubes_clean <- within(youtubes_clean, dislikes[ratings_disabled == "True"] <- dislike_mean);

> mean(youtubes_clean$dislikes);
[1] 2015.779
```

I de forma análoga tractem el camp “comment_count”:

```
> youtubes_clean_no_comments <- youtubes_clean[youtubes_clean$comments_disabled == "False",];
> comments_mean <- mean(youtubes_clean_no_comments$comment_count);
> youtubes_clean <- within(youtubes_clean, comment_count[comments_disabled == "True"] <- comments_m
> mean (youtubes_clean_no_comments$comment_count);
[1] 4910.925
> mean (youtubes_clean$comment_count);
[1] 4910.925
```

Surgeix ara una darrera pregunta. Hem modificat tots els valors a zero? Existeix encara algún registre amb “likes”, “dislikes” o “comment_count” amb valor 0? Comprovem-ho:

```
> youtubes_still_zeros <- youtubes_clean[youtubes_clean$likes == 0 | youtubes_clean$dislikes == 0 | youtubes_clean$comment_count == 0,];

> summary(youtubes_still_zeros);
       video_id     trending_date  category_id        views              likes            dislikes       comment_count    comments_disabled ratings_disabled
 BhIEIO0vaBE:  6   17.16.11:  8   Min.   : 1.00   Min.   :    1141   Min.   :    0.00   Min.   :   0.00   Min.   :   0.0   False:338         False:155
 nx1R-eHSkfM:  6   18.01.02:  7   1st Qu.:17.00   1st Qu.:    6588   1st Qu.:   98.75   1st Qu.:   0.00   1st Qu.:   8.0   True : 62         True :245
 z3XJdMwAvZI:  5   18.05.02:  7   Median :24.00   Median :   54914   Median :38820.00   Median :   0.00   Median : 137.5
 -H90GPnH1q8:  4   18.05.03:  7   Mean   :21.21   Mean   :  961266   Mean   :24083.64   Mean   :  20.39   Mean   :1222.4
 8XiqrERZo_8:  4   18.27.02:  7   3rd Qu.:25.00   3rd Qu.:  240759   3rd Qu.:38820.00   3rd Qu.:   0.00   3rd Qu.:1642.5
 gGnBX6e3RJc:  4   18.30.03:  7   Max.   :29.00   Max.   :51243149   Max.   :38820.00   Max.   :1550.00   Max.   :5953.0
 (Other)    :371   (Other) :357
 video_error_or_removed
 False:397
 True :  3
```

Observem que encara hi ha registres amb dades a zero en els camps “likes”, “dislikes” o “comment_count”. ¿Què hauríem de fer amb aquests registres? ¿Són un nou cas de valor zero que ha sorgit després d’arreglar-ne el primer?

Sense més informació sobre el dataset considero que tinc que tractar aquests valors com a reals malgrat que se’m fa difícil de creure que un video seleccionat entre els 200 de més trending topic en un país no tingui activitat en aquests camps.

Videos marcats com erronis o trets de la xarxa

Ens queda encara per tractar un altre cas de dades que no són ni buïdes ni a zero pero sí incomplertes: els videos que ens venen marcats amb el camp “video_error_or_removed” a cert.

No disposem de informació detallada sobre la casuística d’aquest camp. Hem de suposar que es tracta de videos que contenien algún error de visualització o que han estat trets potser per incomplir les condicions d’us de la plataforma o per decisió del propietari del canal.

Anem a comprovar de quants casos estem parlant:

```
> summary(youtubes_clean);
       video_id     trending_date  category_id        views              likes            dislikes       comment_count    comments_disabled ratings_disabled
 6ZfuNTqbHE8:    8   17.01.12:  200   Min.   : 1.00   Min.   :    1000   Min.   :     0   Min.   :    0.0   Min.   :     0   False:32628       False:32869
 l_lbljSCq0o:    8   17.02.12:  200   1st Qu.:20.00   1st Qu.:  137261   1st Qu.:  2156   1st Qu.:   96.0   1st Qu.:   426   True :  486       True :  245
 UceaB4D0jpo:    8   17.03.12:  200   Median :24.00   Median :  354587   Median :  8926   Median :  298.5   Median :  1350
 9r_rtaye2yY:    7   17.05.12:  200   Mean   :20.75   Mean   : 1113774   Mean   : 38820   Mean   : 2015.8   Mean   :  4911
 Cmg2BnAv4WQ:    7   17.06.12:  200   3rd Qu.:24.00   3rd Qu.:  937967   3rd Qu.: 29280   3rd Qu.:  972.8   3rd Qu.:  3890
 doP7xKdGOKs:    7   17.08.12:  200   Max.   :43.00   Max.   :137843120   Max.   :3014479   Max.   :1602383.0   Max.   :827755
 (Other)    :33069   (Other) :31914
 video_error_or_removed
 False:33088
 True :   26
```

Veiem que hi ha 26 registres amb el camp “video_error_or_removed” a True. Com que són molt pocs casos i tenim poc coneixement de funcionament d’aquest camp decideixo treure’ls de la mostra.

```
> youtubes_clean_no_error <- youtubes_clean[youtubes_clean$video_error_or_removed == "False",];
> summary(youtubes_clean_no_error);
       video_id      trending_date    category_id        views               likes             dislikes       comment_count     comments_disabled ratings_disabl
 6ZfuNTqbHE8:    8   17.01.12:  200   Min.   : 1.00   Min.   :     1000   Min.   :      0   Min.   :      0   Min.   :     0   False:32602       False:32843
 l_lblj8Cq0o:    8   17.02.12:  200   1st Qu.:20.00   1st Qu.:   137248   1st Qu.:   2157   1st Qu.:     96   1st Qu.:   426   True :  486       True :  245
 UceaB4D0jpo:    8   17.03.12:  200   Median :24.00   Median :   354462   Median :   8926   Median :    298   Median :  1350
 9v_rtaye2yY:    7   17.06.12:  200   Mean   :20.75   Mean   :  1113513   Mean   :  38811   Mean   :   2015   Mean   :  4909
 Cmg2BnAv4WQ:    7   17.08.12:  200   3rd Qu.:24.00   3rd Qu.:   937059   3rd Qu.:  29274   3rd Qu.:    972   3rd Qu.:  3889
 doP7xKdGOKs:    7   17.09.12:  200   Max.   :43.00   Max.   :137843120   Max.   :3014479   Max.   :1602383   Max.   :827755
 (Other)    :33043   (Other) :31888
 video_error_or_removed
 False:33088
 True :    0
```

Ara la columna “video_error_or_removed” no aporta informació al conjunt de dades, així que la trec del dataset:

```
> youtubes_clean_no_error <- youtubes_clean_no_error[, c(1,2,3,4,5,6,7,8,9)];
```

No he detectat cap cas més de dades a zero o buïdes.

3.2 Identificació i tractament de valors extrems

Fent una ullada a les dades ja es veu que tenim unes dades molt dispars tot i tractar-se de videos amb les mateixes característiques: tots ells han sigut trending topic.

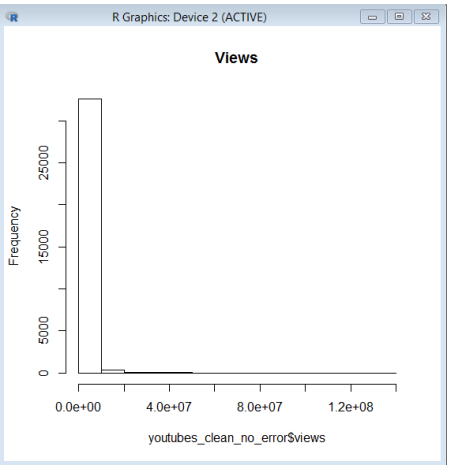
```
     views                likes              dislikes        comment_count
 Min.   :     1000   Min.   :      0   Min.   :      0   Min.   :     0
 1st Qu.:   137248   1st Qu.:   2157   1st Qu.:     92   1st Qu.:   426
 Median :   354462   Median :   8926   Median :    288   Median :  1350
 Mean   :  1113513   Mean   :  38811   Mean   :   2000   Mean   :  4909
 3rd Qu.:   937059   3rd Qu.:  29274   3rd Qu.:    933   3rd Qu.:  3889
 Max.   :137843120   Max.   :3014479   Max.   :1602383   Max.   :827755
```

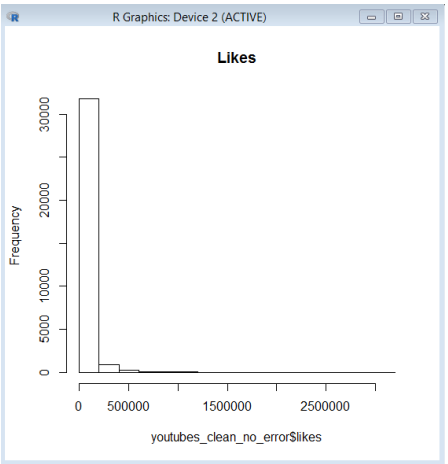
Tenim el mateix comportament en qualsevol de les quatre variables: valors màxims molt per sobre de la mitjana i encara més per sobre del 3er Quadrant.

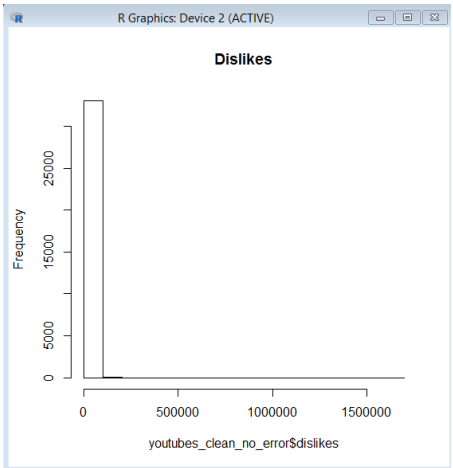
Agafant d’exemple l’atribut “views” veiem que el tercer quadrant ve marcat pel valor 937 mil mentre que el màxim és de més de 137 milions de visites. Per la part inferior tenim també un minim de 1000 que és mil vegades inferior que la mitjana (Mean) del conjunt de la mostra. Podem afirmar doncs que tenim valors extrems.

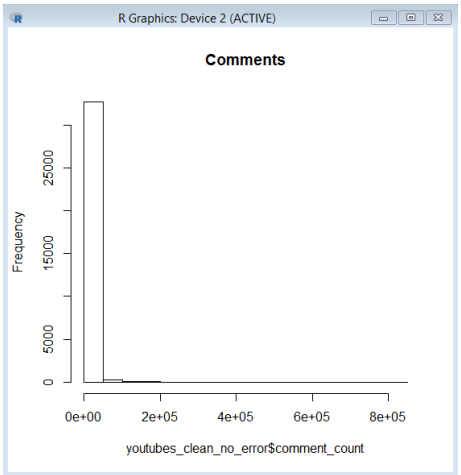
Treuré grafiques per cada atribut per comprovar visualment l’existència de valors extrems i l’impacte que tenen:

```
> hist(youtubes_clean_no_error$views, main="Vistes");
> hist(youtubes_clean_no_error$likes, main="Likes");
> hist(youtubes_clean_no_error$dislikes, main="Dislikes");
> hist(youtubes_clean_no_error$comment_count, main="Comments");
```









Podem observar en les grafiques com l'existència de valors extrems serà un inconvenient en qualsevol estudi estadístic que intentem realitzar. Cap dels histogrames aporta massa informació degut a que els valors extrems provoquen que quasi tota la mostra quedi en una única columna.

Abans de prendre cap mesura pel tractament d'aquests valors extrems hem de intentar esbrinar de quin tipus de outliners es tracta. ¿Són errors en la recollida de dades o són valors reals?

Considero que són valors legítims perquè:

1. Ja hem eliminat de la mostra els registres amb "video_error_or_removed" a cert i, per tant, hem de confiar en que la mostra ve lliure d'errors
2. El propi comportament viral de les xarxes socials afavoreix l'aparició de diferències exponencials entre videos
3. En l'estudi de les dades realitzat fins ara, considero que caldria revisar la metodología de la captura de dades. Crec que agafar una quantitat fixe per dia pot no ser la millor opció ja que hi ha dies de menys activitat que d'altres. Un mètode millor de captura de

dades potser seria analitzar-ho per periodes de temps més amplis (semana, mes, etc) o bé establir uns mínims d'activitat per considerar un video com a trending topic i agafar tots els que cumpleixin aquest mínim encara que en un dia en poguessim recollir 350 i en un altre 50

Per desarrollar més el que exposo en el punt 3, diré que crec que l'estudi té un esbiaixament degut a la manera com les dades han estat recollides pels motius que ja he exposat. Asumiré aquest esbiaixament com a inevitable i segueixo endavant amb l'estudi.

Per tractar valors extrems legítims tenim diferents opcions[1]:

- **Opció 1**: truncar les dades definint un valor màxim i fixar-lo per totes les dades que el superin
- **Opció 2**: transformacions matemàtiques de les dades. Podem, per exemple, aplicar l'arrel quadrada a tots els valors de manera que els més extrems perderan més magnitud que la resta durant el procés.
- **Opció 3**: calcular els estadístics sense tenir en compte els valors extrems
- **Opció 4**: discretitzar els valors i classificar-los segons un conjunt finit de valors com, per exemple: [Molt Baix, Baix, Mitja, Alt, Molt Alt]. Aquesta opció afavoreix aplicar models de classificació
- **Opció 5**: normalitzar tots els valors al interval [0,1]. Aquesta opció afavoreix l'estudi gràfic de les dades

No existeix una opció correcta universal, hem de tenir en compte l'objectiu de cada análisis. En el nostre cas l'objectiu és fer un estudi de l'activitat que reben els videos més vistos en funció de la categoria per identificar-ne les més interessants a l'hora de penjar-hi contingut professionalment.

No ens interessen les dades concretes, sino analitzar conjunts d'èxit. Per una banda, la quantitat de videos que conté cada categoria ja és un indicador perquè recordem que el conjunt de dades ja conté videos trending topic. A part d'això, ens interessa tenir mesures de videos de molt èxit, d'aquesta manera podrem analitzar cada categoria des de dues perspectives: volum de videos i qualitat del volum (si conté molts videos de molt èxit o no).

Per tant considero que ens interessa aplicar transformacions que discretitzin els valors en {Baix, Mitja, Alt }.

Les comandes R per discretitzar aquests valors són:

Explicaré al detall cada pas per la columna "Views" i després inclouré les comandes per la resta d'atributs. He creat un alias "yf" al dataset per reduir el tamany de les comandes.

Afegeixo el camp discretit per les vistes amb el seu valor per defecte:

```
> yf$d_views="Mitja";
```

[1] Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369

Actualitzo els valors per sota del primer quadrant amb el valor “Baix”:

```
> # lower quad
> yf <- within(yf, d_views[yf$views <= 137248 ] <- "Baix");
```

El valor 137248 està tret de fer un summary(yf) i prenent el valor proporcionat per “1st Qu.”

Actualitzo els valors per sobre del tercer quadrant amb el valor “Alt”:

```
> # upper quad
> yf <- within(yf, d_views[yf$views >= 937059 ] <- "Alt");
```

Aconseguim així un distribució dels valors en {Baix, Mitja, Alt} amb significació estadística.

Repetim els procés per la resta de variables:

```
> yf$d_likes="Mitja";
> # lower quad
> yf <- within(yf, d_likes[yf$likes <= 2059 ] <- "Baix");
> # upper quad
> yf <- within(yf, d_likes[yf$likes >= 28189 ] <- "Alt");
> yf$d_dislikes="Mitja";
> # lower quad
> yf <- within(yf, d_dislikes[yf$dislikes <= 92 ] <- "Baix");
> # upper quad
> yf <- within(yf, d_dislikes[yf$dislikes >= 933 ] <- "Alt");
> yf$d_comment_count="Mitja";
> # lower quad
> yf <- within(yf, d_comment_count[yf$comment_count <= 426 ] <- "Baix");
> # upper quad
> yf <- within(yf, d_comment_count[yf$comment_count >= 3889 ] <- "Alt");
```

4. Anàlisi de les dades

4.1 Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar)

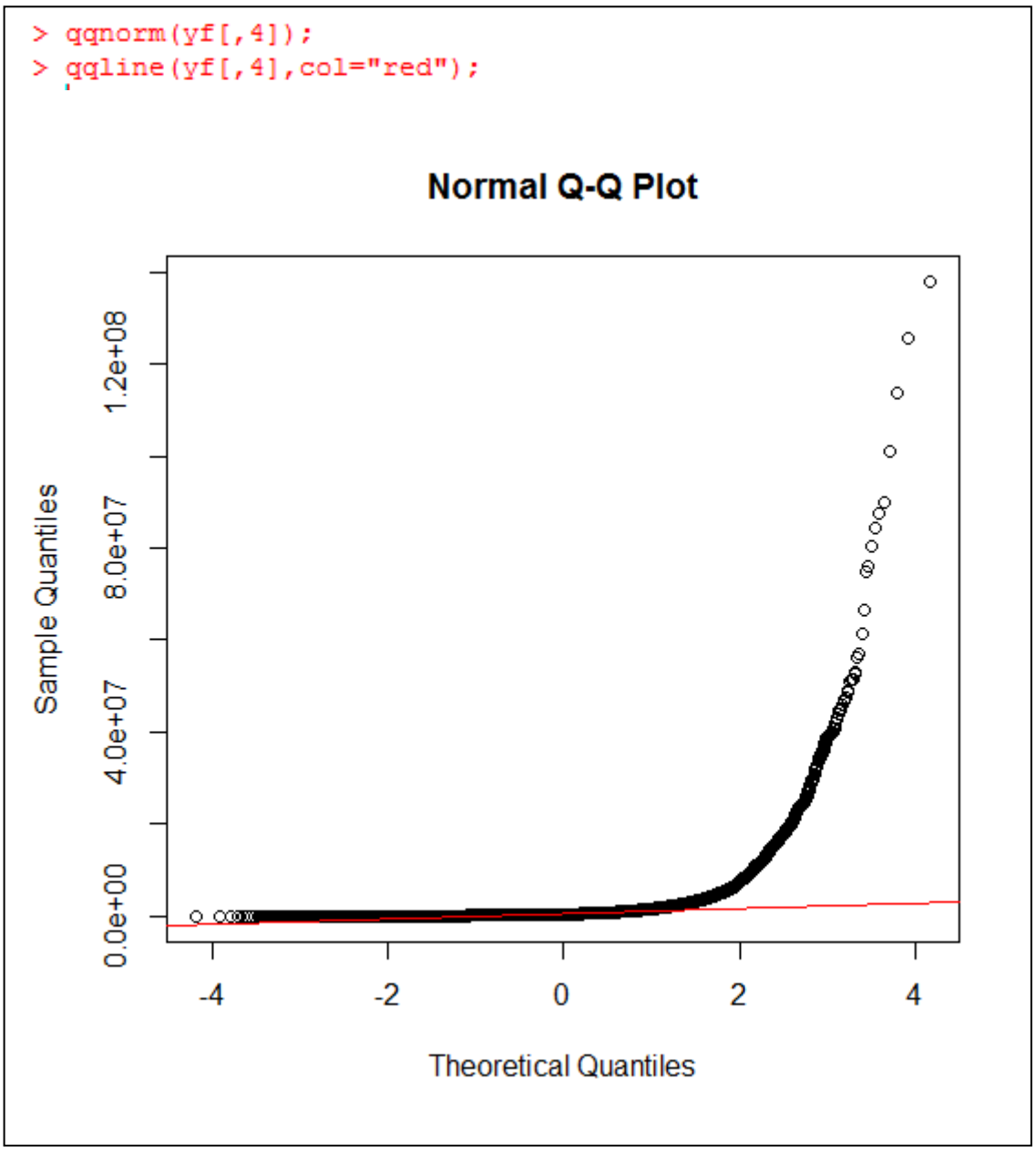
Tenim les nostres dades discretitzades però diferents mesures a triar. Hauríem d’analitzar per número de visites? Per “likes”? Per nombre de comentaris? Sospito que el nombre de visites ja es un indicador suficient ja que per lògica a més visites més probalitats d’aconseguir algún “like” o de que algú hi faci algún comentari. Comprovarem aquesta hipòtesi buscant correlació entre vistes i la resta de mesures. Un cop demostrada aquesta correlació podrem extreure les nostres conclusions analitzant la relació entre el nombre de visites i la categoria a la que pertany el video.

4.2 Comprovació de la normalitat i homogeneïtat de la variancia

Comprovarem la normalitat de cadascuna de les variables numèriques implicades en el dataset mitjançant la representació visual per Q-Q plot. En dibuixarem la distribució dels punts i la linea teórica de normalització. Si la majoria de punts segueixen la linea llavors podem afirmar que segueixen una distribució normal.

Vistes

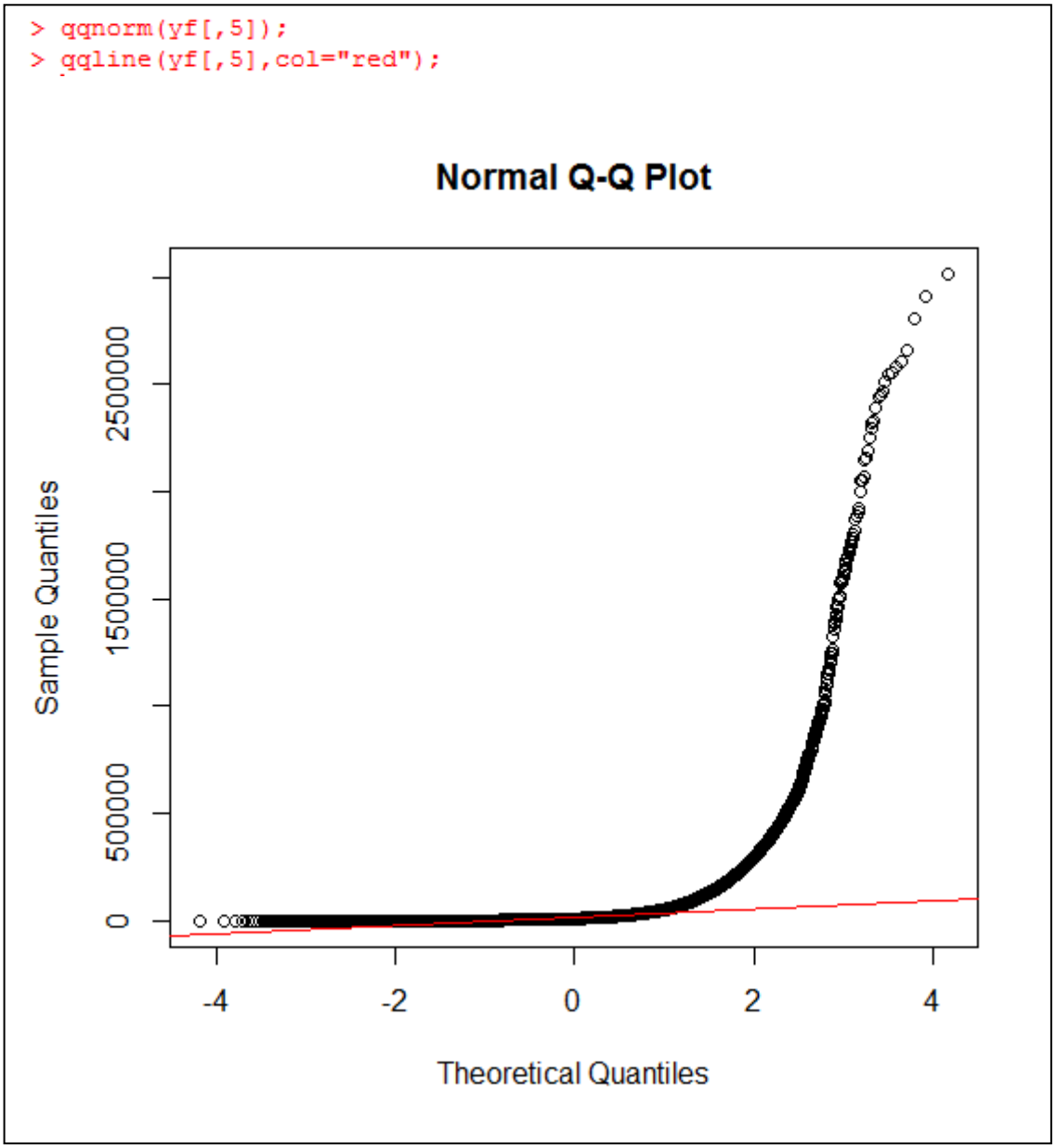
```
> qqnorm(yf[,4]);
> qqline(yf[,4],col="red");
```



Podem veure l'efecte de no haver tret els outliners de les dades originals. Tot i això considero que està provat que segueixen una distribució normal.

Likes

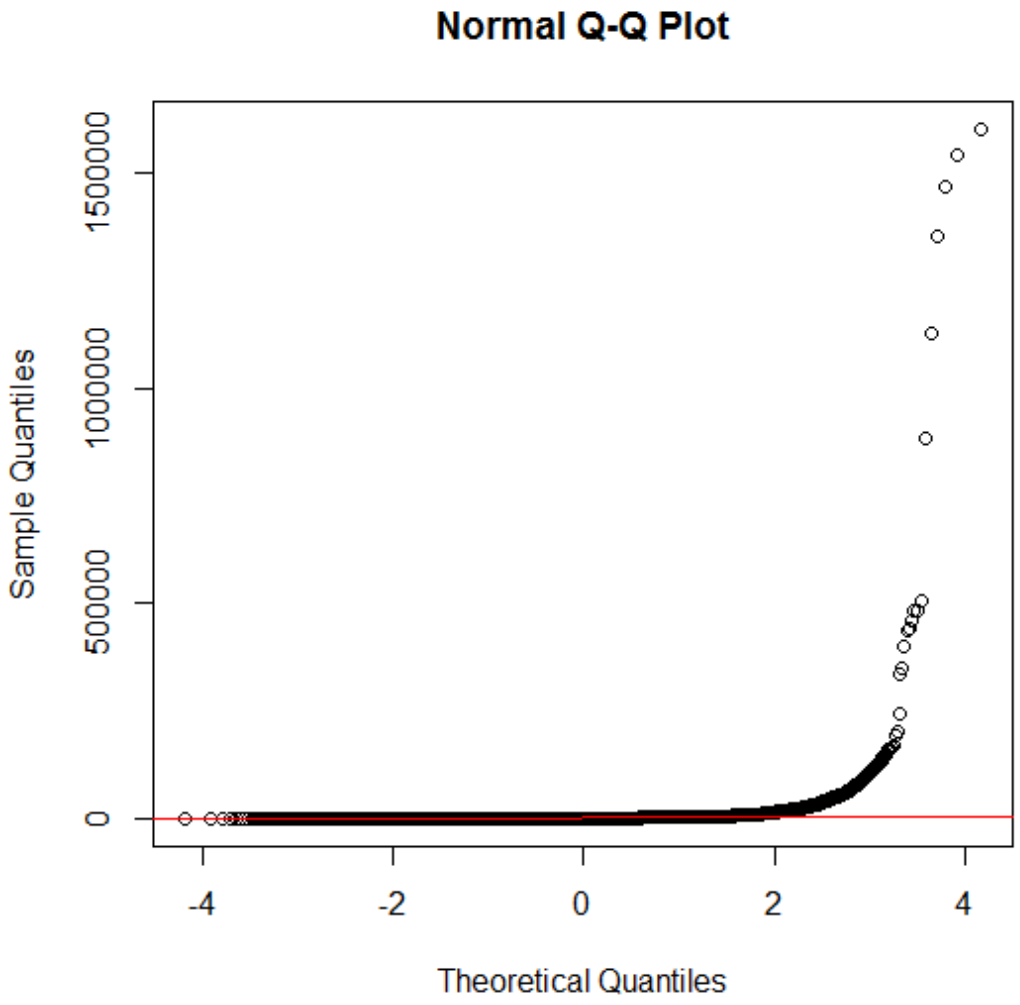
```
> qqnorm(yf[,5]);
> qqline(yf[,5],col="red");
```



Observem el mateix comportament que amb les vistes.

Dislikes

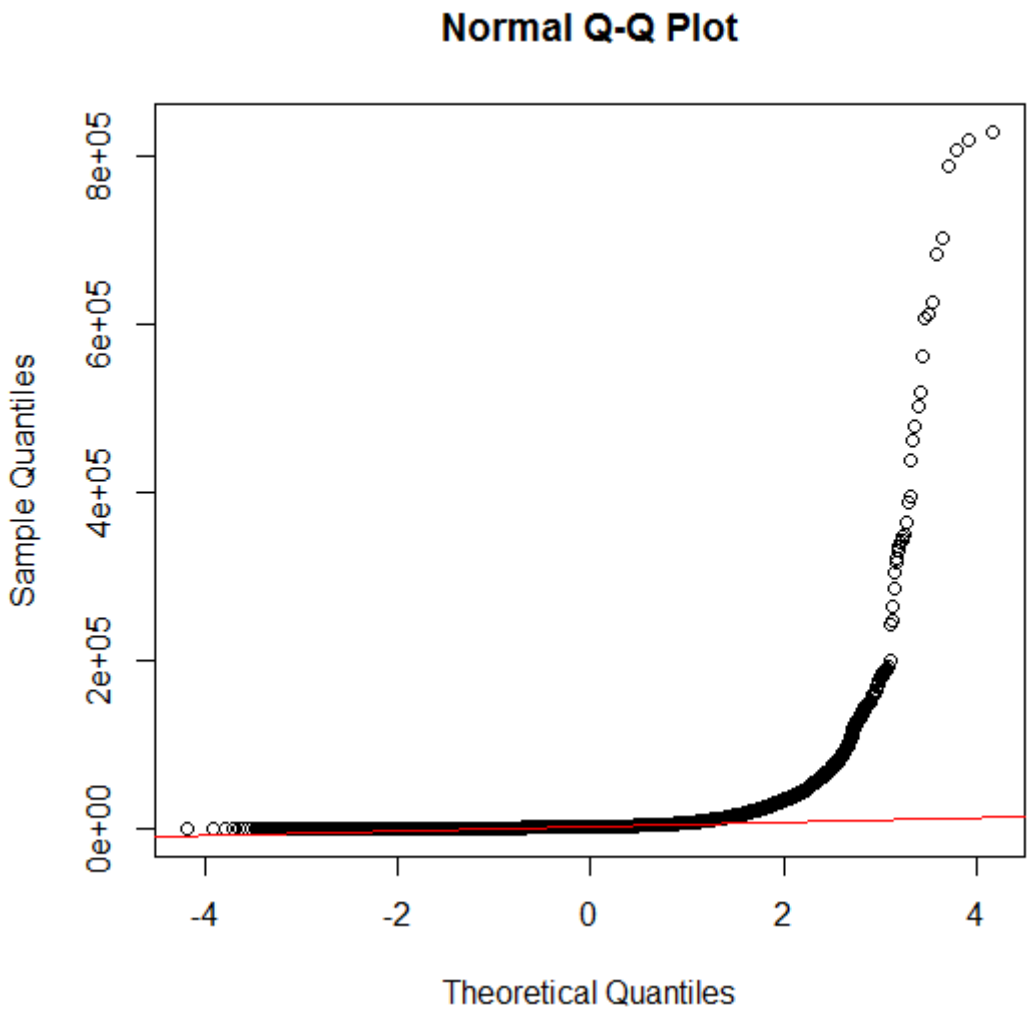
```
> qqnorm(yf[,6]);
> qqline(yf[,6],col="red");
```



Observem el mateix comportament que amb les vistes.

Comment count

```
> qqnorm(yf[,7]);
> qqline(yf[,7],col="red");
```



Observem el mateix comportament que amb les vistes.

4.3 Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

Les grafiques anteriors quasi idèntiques ja ens indiquien que hi haurà una correlació entre variables. Anem a verificar-ho amb proves estadístiques.

Utilitzarem el test de correlació de Pearson el qual ens dona un valor entre 1 i -1 que indica la relació entre dues variables i on 1 significa una correlació positiva absoluta i -1 negativa. El valor 0 indicaria que les dues variables no tenen cap correlació.

Correlació Visites – Likes

```
> cor.test(yf$views,yf$likes,method="pearson");

        Pearson's product-moment correlation

data:  yf$views and yf$likes
t = 269.26, df = 33086, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8252370 0.8319902
sample estimates:
      cor
0.8286438
```

El valor de correlació és de 0.8286438 i indica una correlació positiva molt alta. Podem afirmar que a més visites, més “likes”.

Correlació Visites – dislikes

```
> cor.test(yf$views,yf$dislikes,method="pearson");

        Pearson's product-moment correlation

data:  yf$views and yf$dislikes
t = 124.47, df = 33086, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5573462 0.5720240
sample estimates:
      cor
0.5647297
```

Curiosament observem un valor molt diferent quan analitzem la correlació entre les visites i el número de "dislikes". Tot i que continua sent un valor positiu (a més vistes, més "dislikes"), aquest marca una relació molt menys marcada.

Una possible explicació a aquest fenòmen seria que costa més als usuaris penalitzar negativament un video. La reacció més habitual quan un video no ens agrada és veure'n un altre o tancar el navegador i, per tant, aquesta mesura deu tenir molts menys "dislikes" dels que realment són i d'aquí que no aumenti tant positivament en funció del nombre de vistes.

Correlació vistes – numero de comentaris

```
> cor.test(yf$views,yf$comment_count,method="pearson");

        Pearson's product-moment correlation

data:  yf$views and yf$comment_count
t = 172.11, df = 33086, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6815605 0.6929316
sample estimates:
      cor
0.6872882
```

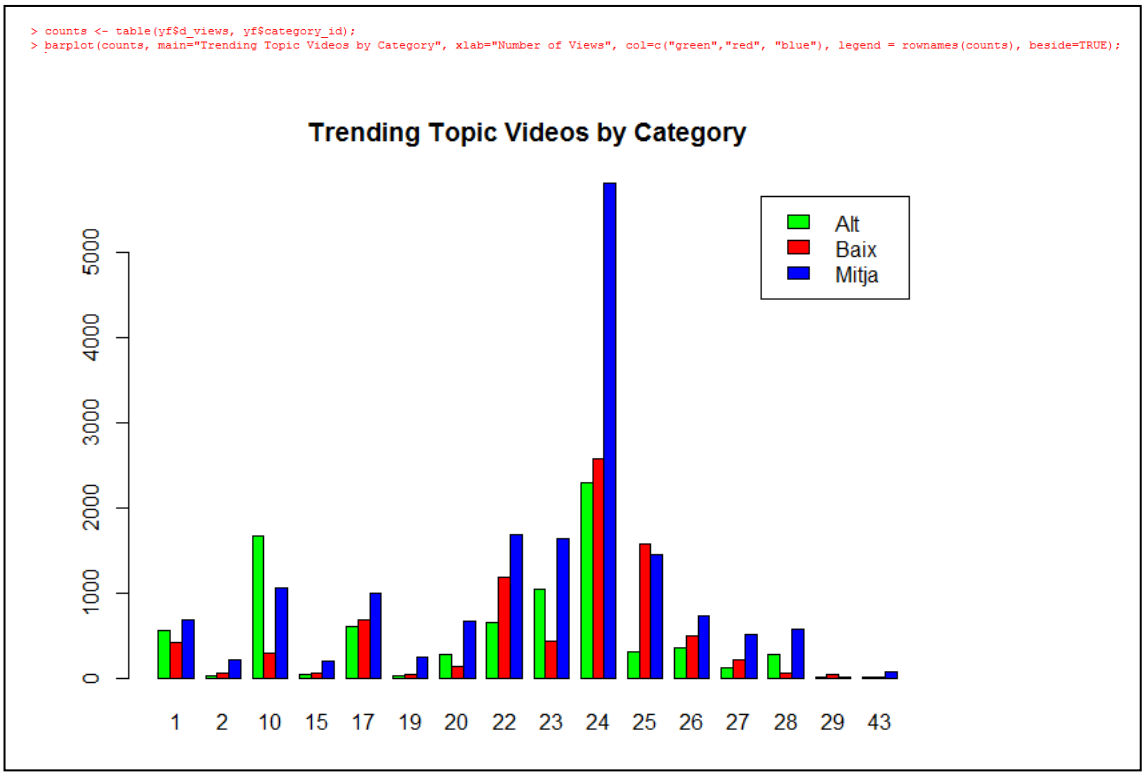
Observem un valor positiu prou alt (proper a 1) i, per tant, podem afirmar que existeix relació entre les visites i el número de comentaris.

Podem concloure doncs que per analitzar l'activitat generada en els videos la mesura de les vistes es un indicador fiable del nostre conjunt de dades.

5. Representació dels resultats a partir de taules i grafiques

Per analitzar el comportament dels videos trending topic en funció de la categoria utilitzare un barplot:

```
> counts <- table(yf$d_views, yf$category_id);
> barplot(counts, main="Trending Topic Videos by Category", xlab="Number of Views", col=c("green","red", "blue"), legend = rownames(counts), beside=TRUE);
```



Recordem que estem fent servir la columna discretitzada “d_views” que ens agrupa els videos segons un numero alt, mitjà o baix de visites. Les categories estan representades per un codi intern present a les dades. La següent taula mostra els noms de les categories presents en el gràfic segons la documentació adjunta amb el dataset:

Id	Nom
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	<descripció no disponible>
43	Shows

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Estem en disposició de donar respostes a les preguntes plantejades.

¿És Youtube una plataforma principalment de música? Si presuposem que els videos que arriben a ser trending topic són una representació proporcional del total de videos existents a la xarxa està clar que no. La categoria de música té poca representació en la gràfica.

¿Existeix una relació entre el nombre de visites i el número de likes, dislikes i número de comentaris? Sí, tal i com cabia esperar. Hem descobert que la relació no és tan marcada en quant al número de "dislikes", fet que sense més dades atribueixo a que ens costa més penalitzar un video amb un "no m'agrada" que no pas deixar-hi una valoració positiva.

¿En quina categoria ens hauríem d'especialitzar si volguessim dedicar-nos professionalment a la pujada de videos? Aquesta pregunta és més complexe de respondre ja que dependrà de la nostra estrategia com a youtubers. Si ens basem només en el volum total de videos que hi apareixen ens hauríem de decantar per l'entreteniment en concret, o be en contingut tipus "blog" en general (les categories "People & Blogs", "Comedy", "Entertainment" i "News & Politics" representen la major part de la mostra).

En canvi,si volguessim apostar per aconseguir videos concrets amb un alt nivell de visites hauríem d'arriscar-nos amb la música o l'entreteniment, ja que són les dues categories que més proporció de franja "Alt" tenen de la mostra.

A partir de la distribució de les franges "Alt", "Mitja" i "Baix" intrepreto que les categories amb un comportament més normalitzat, és a dir, amb una presència majoritària de franja mitjana i equilibri entre franges alta i baixa, són "Sports", "People & Blogs" i "Entertainment". Crec que això pot ser un indicador de que tenen una presència més sostinguda entre els trending topics a diferència de, per exemple, la categoría de música que al tenir una altíssima proporció de la franja alta i mitjana fa pensar que és deu més a fenòmens virals espontànis.

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python

El codi fet servir durant aquesta pràctica està disponible al repository de github creat expressament per la mateixa.