Análisis de Estadísticas de Videos en Tendencia de YouTube

Integrantes:

- Caparachin Villaverde, Nanto Gustavo (E202110812)
- García Godos Villavicencio, Jorge Daniel (E202110981)
- Huamani Franco, Ismael (E202110991)
- Poma Ludeña, Rodrigo Renato (E202111018)

INTRODUCCIÓN

Extracción e Ingesta de Información

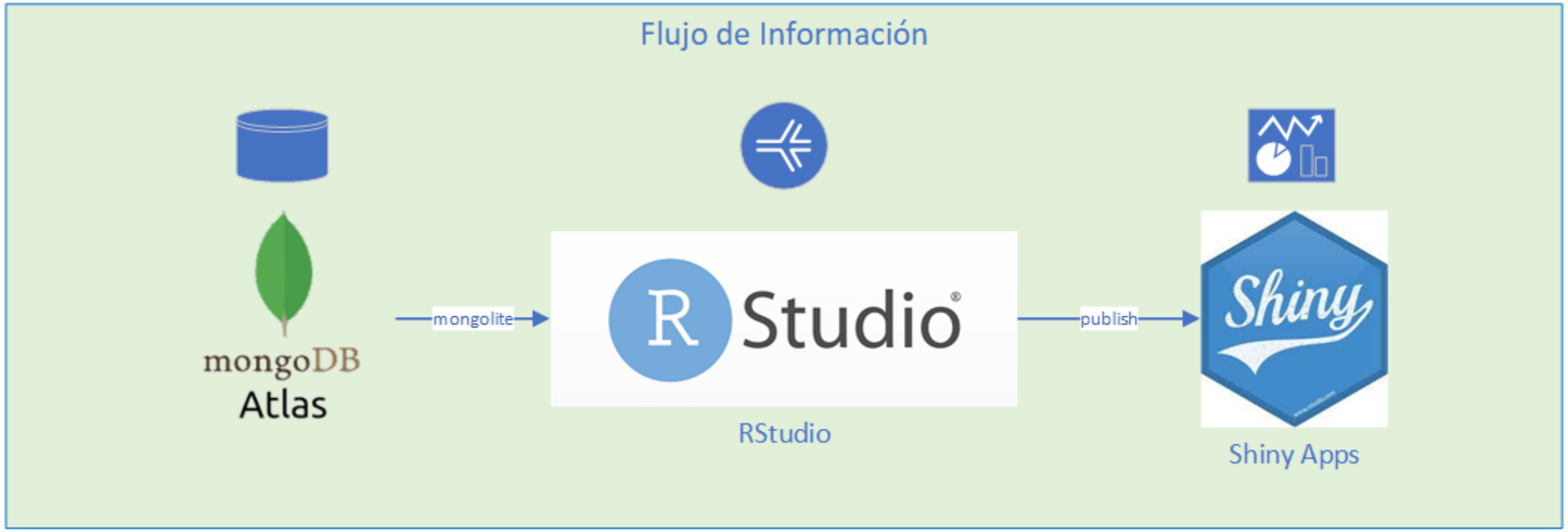
Fuente:
Trending YouTube Video Statistics

Agente:
Importador de archivos hacia MONGO DB (ATLAS).

Base de Datos:
Servicio MongoDB alojado en la nube en **AWS**, Azure y Google Cloud.

INTRODUCCIÓN

Extracción, Análisis y Representación de los Datos



Fuente:
MongoDB con las colecciones de datos unificados

Aplicación de Análisis:
Entorno de desarrollo integrado con herramientas de análisis estadístico.

Herramienta de Visualización:
Aplicaciones interactivas de visualización de datos

METODOLOGÍA

Procedimiento realizado

1. Identificación de fuente de datos
   Buscamos fuentes de datos confiables y escogimos una fuente considerando la veracidad, volumen y variedad.
2. Dimensionamiento y creación de la base de datos
   Dimensionamos la base de datos considerando solo tres países y creamos la base de datos en su entorno PAAS creando usuarios de solo lectura y abriendo la comunicación a cualquier IP.
3. Obtención de datos e importación a base de datos
   Automatizamos la obtención de datos usando un script bash que descarga las fuentes de datos y los ingresa en la base de datos mongo con la librería de mongo-import.
4. Conexión a base de datos y extracción de data
   Utilizamos mongolite para establecer la conexión con la base de datos y extraer la información.
5. Comprensión y limpieza de datos
   Organizamos, categorizamos, tipificamos y unificamos las diferentes colecciones para tener un dataframe confiable en el que realizar nuestro análisis.
6. Análisis de outlines
   Normalizamos los datos, aplicamos diferentes técnicas como Shapiro, Anderson-Darling, Mahalanobis y Chi-Cuadrado
7. Exploración de datos en Shiny
   Diseñamos los componentes UI y completamos diferentes gráficos de hallazgos en la sección se server.

Extracción de los datos CSV & JSON

Cadena de Conexión a MongoDB desde RStudio

Extracción de Colecciones

Tratar los datos extraídos según el tipo de fuente asociado (CSV, JSON).

```
#1.1 Coneccion a la base de datos
connection_string = 'mongodb+srv://read-only-user:U5PN6pQ1eirrSu7e@mycluster.zal73.mongodb.net/dbyoutube'

#1.2 Extrayendo la data de cada coleccion tipo CSV de trending
MX_trending = mongo(collection="MX_youtube_trending_data", db="dbyoutube", url=connection_string)
US_trending = mongo(collection="US_youtube_trending_data", db="dbyoutube", url=connection_string)
FR_trending = mongo(collection="FR_youtube_trending_data", db="dbyoutube", url=connection_string)

#1.3 Eplorando las colecciones de la base de datos de YOUTUBE
MX_trending$run('{"listCollections": 1}')$cursor$firstBatch$name

#1.4 Extrayendo la data de cada coleccion tipo JSON de category
MX_category = mongo(collection="MX_category_id", db="dbyoutube", url=connection_string)
US_category = mongo(collection="US_category_id", db="dbyoutube", url=connection_string)
FR_category = mongo(collection="FR_category_id", db="dbyoutube", url=connection_string)

#1.5 Armamos los archivos de trending que son tipo csv
MX <- MX_trending$find('{}')
US <- US_trending$find('{}')
FR <- FR_trending$find('{}')

#1.6 Armamos los archivos de categoria que son tipo json
cat_MX <- MX_category$find('{}')
cat_US <- US_category$find('{}')
cat_FR <- FR_category$find('{}')
```

Exploración de los datos CSV

Reconocimiento de los datos en los 5 primeros registros

Identificación de una columna de tags con potencial de análisis

Hallazgo de datos propios de un video en una plataforma digital

```
> head(MX)
     video_id          publishedAt    channelTitle categoryId        trending_date
1 QgX0zn5nMyk 2020-08-11T15:57:47Z             Hoy         24 2020-08-12T00:00:00Z
2 zU5lgsAKIIA 2020-08-11T20:18:23Z           MaiLu         22 2020-08-12T00:00:00Z
3 p7WtYeowOEY 2020-08-12T01:00:04Z SOY TRIKI TRIKI         22 2020-08-12T00:00:00Z
4 M9Pmf9AB4Mo 2020-08-11T17:00:10Z    Apex Legends         20 2020-08-12T00:00:00Z
5 eew7QJqFn7Q 2020-08-10T17:06:34Z        Rubius Z         20 2020-08-12T00:00:00Z
6 vX5v6umIWXU 2020-08-10T16:59:41Z          Genius         10 2020-08-12T00:00:00Z




  tags
1                                                          Televisa|Televisa espectaculos|programa hoy televisa|Pr
ograma Hoy|elenco programa hoy|videos programa hoy|programa hoy 2020|Andrea Legarreta|Galilea Montijo|Raúl Arai
za|Andrea Escalona|#ConLasEstrellas|Paul Stanley|Elenco programa hoy 2020|Lambda García|Marisol González|#Telev
isaTeAcompaña|Magda Rodríguez|con permiso|unicable|pepillo origel|martha figueroa|erika buenfil|hijo de erika b
uenfil
2                                                                                  lulu99|novio de lulu99|lulu y
 su novio|maicol y luisa|tag del novio|novio|lulu 99|99|lulu99 y su novio canal|mailu|mai lu|pareja|novios|YOLO
 aventuras|sandra cires|24 horas|esposados|con mi novio|maicol|maiking|¡NUEVO CANAL JUNTOS! Lulu Y Maicol <U+27
64><U+FE0F> NUESTRO PRIMER VIDEO!!|novio vs novia|MaiLu|reto|quien conoce mejor al otro|challenge|skabeche|123
 spanish
3


                                                                                 Triki triki|Payaso triki trki|Soy triki
 triki
4 Apex Legends|Apex Legends characters|new Apex Legend|Apex Legends Rampart|Apex Legends Season 6|Apex Legends
 Boosted|Battle Pass|Season 6 Battle Pass|Apex Legends new season|Apex Legends game|Respawn Apex Legends|Battle
 Royale game|Battle Royale|Battle Royale shooter|Apex Games|squad play|multiplayer shooter|Apex Legends PS4|Ape
x Legends Xbox|Apex Legends PC|Apex Legends Origin|Respawn Entertainment|Electronic Arts|Titanfall 2|fun battle
 royale
5

                                                               elrubius|rubius|kun|kun aguero|agüero|fall guys|fall|g
uys|full|battle royale|gameplay|mejor|mejores momentos|manchester|manchester city|futbol|2020|gol|goles|sergio|
kubius
6

            genius|rap genius|verified|official lyrics|lyrics|lyric video|Lyric videos|pop music|hip hop|rap|ne
w pop music|jd pantoja vevo|jd pantoja hagamos las paces|jd pantoja canciones|jd pantoja letra|jd pantoja 12 19
|gbwc0
  view_count  likes dislikes comment_count comments_disabled ratings_disabled country
1     521751   4680     3129          2285             False            False  Mexico
2     989033 157522     2060         18314             False            False  Mexico
3      64858   6327      110           437             False            False  Mexico
4    2381688 146744     2794         16557             False            False  Mexico
5    4331471 492848     4011         10942             False            False  Mexico
6     403952  37242     9986          6734             False            False  Mexico
```

Limpieza de los datos CSV

Creamos una variable para identificar los países

Eliminamos aquellas variables que no se considerados evaluar, tal como:
Link del video,
Descripción,
Titulo,
El código del canal.

```
#2.1.1 Preparamos las tablas de YOUTUBE creando un nuevo campo country: Mexico,
MX$country = MX$country = 'Mexico'
US$country = US$country = 'USA'
FR$country = FR$country = 'Francia'


#2.1.2. Eliminamos las columnas que no vamos a utilizar
MX[, c("thumbnail_link", "description", "title", "channelId")] <- NULL #tags ar
US[, c("thumbnail_link", "description", "title", "channelId")] <- NULL
FR[, c("thumbnail_link", "description", "title", "channelId")] <- NULL
#se elimina thumbnail_link porque es la direccion url no agrega valor al anali
#se elimina description porque es la descripcion del video, contiene datos no e
```

Solo en la variable de canal (CHANNELID) se encontraron vacíos, además, no brinda un valor significativo al análisis, por lo que, no se utilizará en la evaluación

Procesamiento de los archivos JSON

Encontramos los datos de interés de los archivos JSON,

Creamos la función para extraer los datos de interés: código de categoría y descripción de la categoría

```
cat_FR$snippet$title[1]                                    #
cbind(id=cat_FR$id[1], title=cat_FR$snippet$title[1])      #
```

```
> cat_FR$snippet$title[1]
[1] "Short Movies"
> cbind(id=cat_FR$id[1], title=cat_FR$snippet$title[1])
     id   title
[1,] "18" "Short Movies"
>
```

```
#Inf de categoria de Mexico
vector1 = c()
vector2 = c()
for (i in 1:length(cat_MX$id)){
  vector1 = c(vector1, cat_MX$id[i])
  vector2 = c(vector2, cat_MX$snippet$title[i])
}
DFX = data.frame("id"=vector1, "title"=vector2)

#Inf de categoria de USA
vector3 = c()
vector4 = c()
for (i in 1:length(cat_US$id)){
  vector3 = c(vector3, cat_US$id[i])
  vector4 = c(vector4, cat_US$snippet$title[i])
}
DFUS = data.frame("id"=vector3, "title"=vector4)

#Inf de categoria de Francia
vector5 = c()
vector6 = c()
for (i in 1:length(cat_FR$id)){
  vector5 = c(vector1, cat_FR$id[i])
  vector6 = c(vector2, cat_FR$snippet$title[i])
}
DFFR = data.frame("id"=vector5, "title"=vector6)

#2.2.3. juntamos las 3 bases de datos
DFCAT = bind_rows(DFX,DFUS,DFFR)
```

Limpieza final, unión y conversión de los tipos de datos

Eliminación de duplicados

Preparación para la unión

Unión a la izquierda

Conversión a Factor de los registros de Categoría

```
#2.2.4 Revisamos y limpiamos los datos antes de aplicar la union

DFCATEGORIA = DFCAT[!duplicated(DFCAT), ] #unique(DFCATEGORIA[c("title")]) --- se revisa que no haya dat

str(DFCATEGORIA) #observamos que ID esta en valor STR
str(YTB)         #observamos que category_id esta como NUM, lo mejor sera realizar su modificacion

YTB$categoryId = as.integer(YTB$categoryId) #realizamos la conversion de num a str
table(YTB$categoryId)                        #revisamos los datos sobre los que vamos a preparar la unio

colnames(DFCATEGORIA) = c("categoryId","category") #para aplicar el merge, lo mejor es darle renombrar y

#2.3 YOUTUBE: Creacion de la base de datos final para trabajo

#2.3.1. realizamos el merge considerando los valores que estan a la izquierda
YOUTUBE = merge(x = YTB, y = DFCATEGORIA, all.x = TRUE)

#2.3.2. volvemos a revisar y hacer una limpieza final
YTB$category <- as.factor(YTB$category);  #vemos que falto aplicar factor a categoria
levels(YTB$category) <- c("Film & Animation", "Autos & Vehicles",          "Music",
                          "Pets & Animals"  ,  "Sports"         ,          "Travel & Events",
                          "Gaming"          , "People & Blogs"  ,          "Comedy",
                          "Entertainment"   , "News & Politics" ,          "Howto & Style",
                          "Education"       ,     "Science & Technology" , "Nonprofits & Activism",
                          "Movies"          ,      "Shows"   ,             "Trailers"
)
```

METODOLOGÍA

Herramientas utilizadas

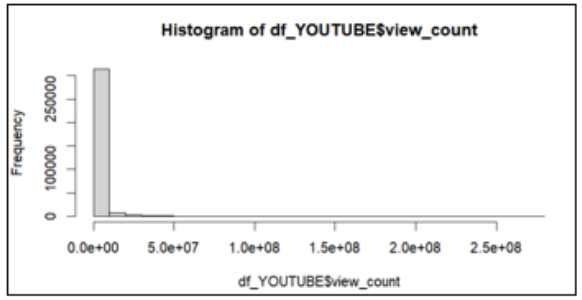
HERRAMIENTA	DETALLE	UTILIZACIÓN
Kaggle	Kaggle es un repositorio de datasets abiertos.	Obtención de nuestro dataset actualizado de tendencias de videos.
Amazon Cloud9	IDE de desarrollo de software ejecutado en EC2.	Manipulación del comprimido del dataset e importación a la base de datos.
Atlas MongoDB	Base de datos no relacional administrada por Atlas.	Base de datos de consulta de dataset.
DataBase Tools	Librería de manipulación de datos	Uso de la función mongo-import para la importación del dataset.
RStudio Cloud	IDE cloud de desarrollo de proyectos en lenguaje R.	Desarrollo de análisis de datos.
Amazon EC2	Máquina virtual multipropósito.	Ejecución de consultas con alto consumo de datos y procesamiento.

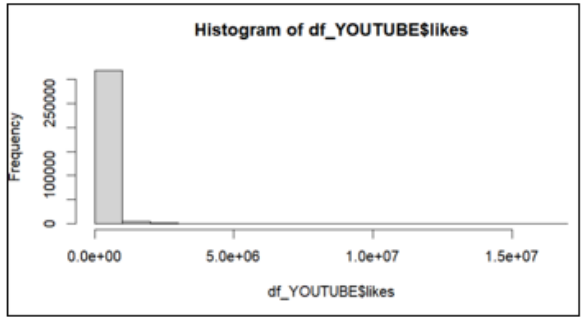
METODOLOGÍA

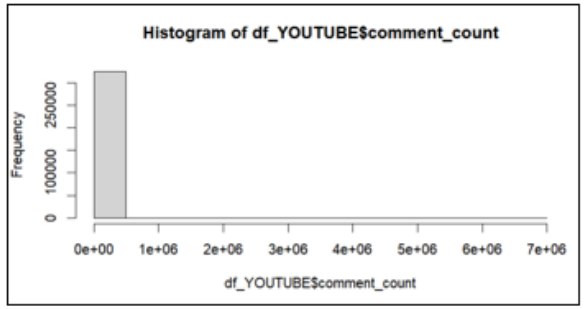
TÉCNICA	ETAPA DE UTILIZACIÓN
Eliminación de duplicados	Comprensión y limpieza de datos
Aplicación de Factor o Levels	Comprensión y limpieza de datos
Validación de Nulos	Comprensión y limpieza de datos ,Análisis de outliers
Validación de Ceros	Análisis de outlines
Normalización usando criterio Min-Max	Análisis de outlines
Shapiro	Análisis de outlines
Anderson-Darling	Análisis de outlines
Distancia de Mahalanobis	Análisis de outlines, Exploración de datos en Shiny
Chi cuadrado	Análisis de outlines, Exploración de datos en Shiny
Análisis de frecuencia	Exploración de datos en Shiny
Análisis de dispersión	Exploración de datos en Shiny

RESULTADOS

Análisis de dispersión







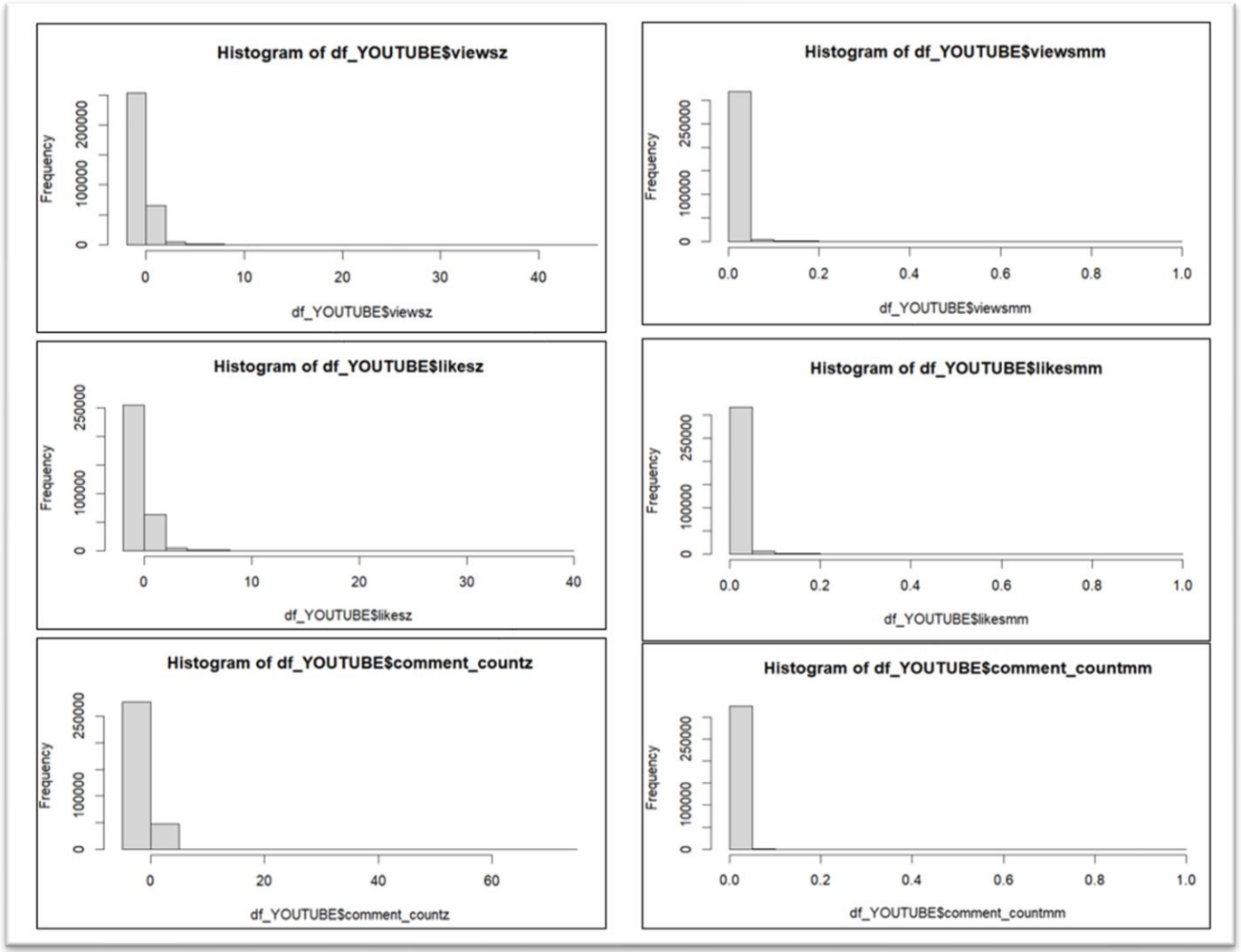
Resultado 1:
Inicialmente utilizamos 4 variables para el análisis: Likes, Views y Coments y dislikes hemos encontrado que la distribución de las variables de Likes, Views y Coments, tienen una distribución claramente asimétrica positiva.

RESULTADOS

Normalización

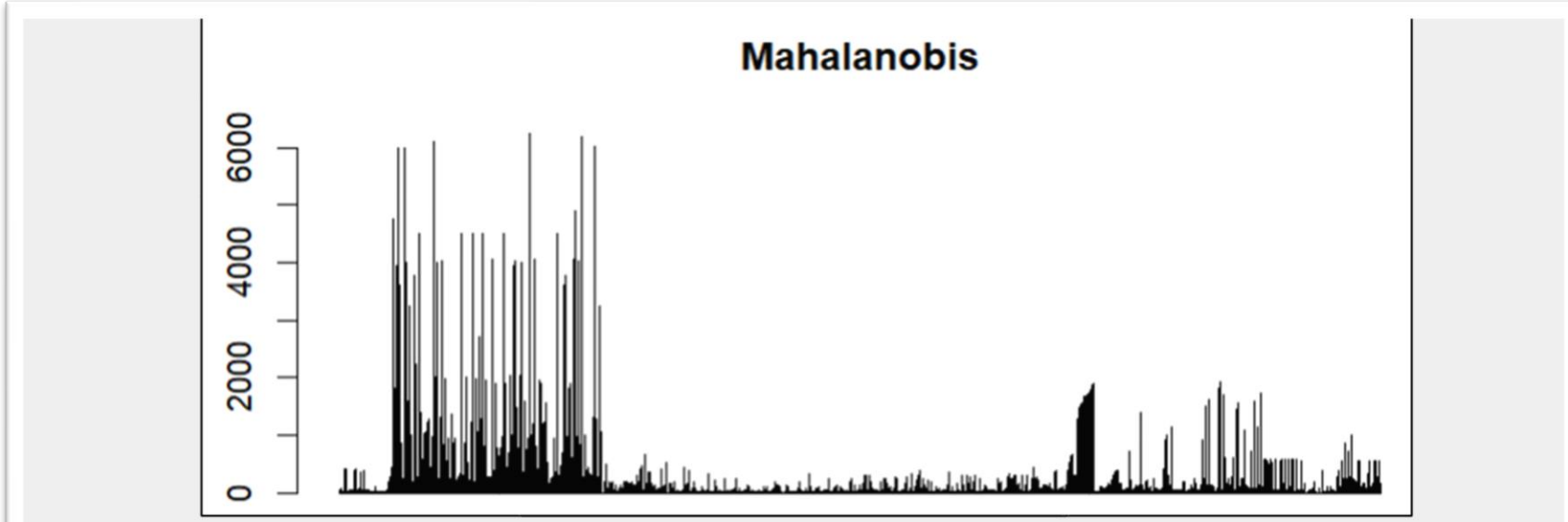
Resultado 2:
Al ser los rangos en los cuales se distribuye cada variable son muy diferentes por lo que se procedió a normalizarlos, tanto por el método de Puntaje Z, para que conserve la distribución de los resultados, como MINMAX.

- Método de Puntaje Z
- Método de MINMAX



RESULTADOS

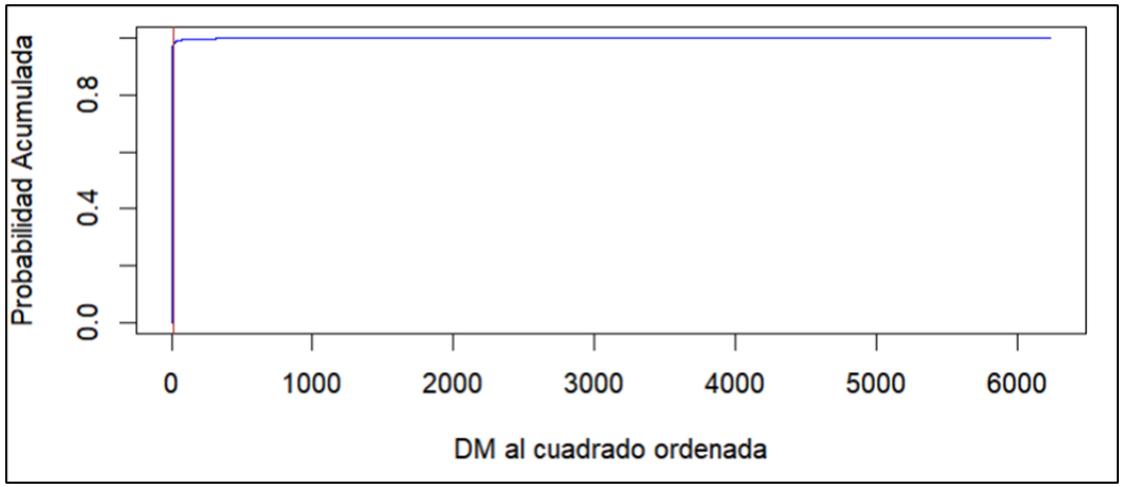
Resultado 3:
Hemos logrado identificar aquellos valores que en su combinación de sus variables views, likes y coments presentan un comportamiento muy diferenciado al resto, logrando valores lejos del centro alcanzando un valor de 6,241.



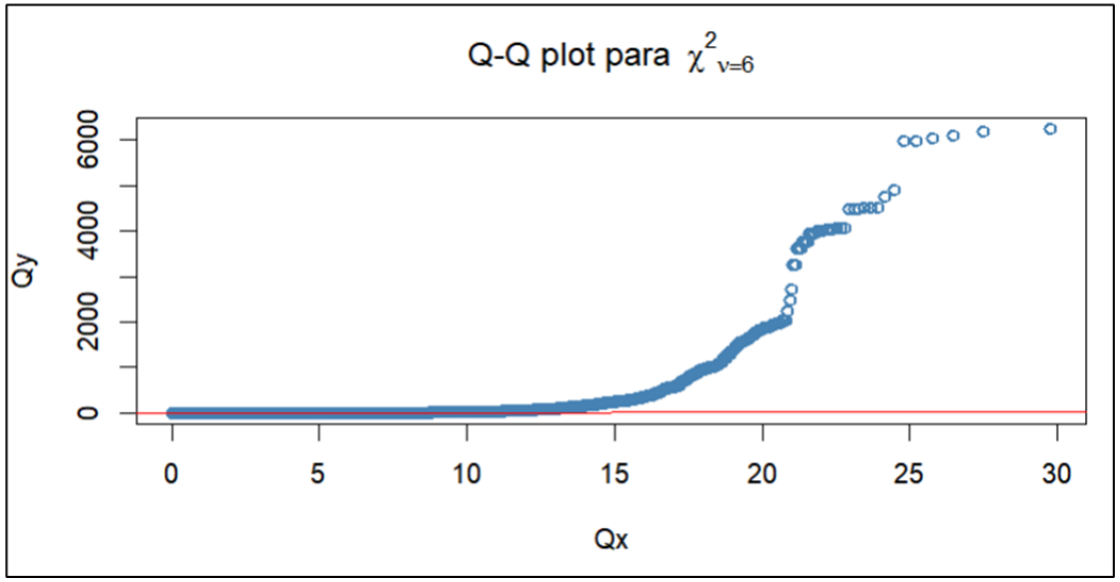

```
dmX[order(dmX, decreasing=TRUE)]      # Valores ordenados
 [1] 6241.9481 6174.6837 6087.5568 6026.6682 5974.7298 5974.6639 4904.8350 4755.1686 4501.4315
[10] 4501.3477 4501.2585 4491.3957 4491.3852 4491.3624 4052.6728 4052.6044 4052.5759 4025.7232
[19] 4025.5937 4025.4940 3992.9173 3992.8885 3992.8532 3941.1282 3941.1132 3941.0153 3762.3705
[28] 3762.3158 3762.1394 3609.6299 3609.5634 3609.4146 3248.5931 3248.5884 2706.3610 2467.6685
[37] 2223.4181 2037.2953 2037.2560 2014.1696 2014.1562 1978.9732 1978.9282 1978.9125 1943.1308
[46] 1942.9957 1942.9774 1914.9581 1904.5193 1891.6663 1891.4509 1891.4359 1890.0737 1866.0360
[55] 1847.0950 1826.4574 1826.4227 1826.1002 1824.4694 1798.9475 1779.3315 1763.5663 1745.0841
[64] 1726.4912 1711.3655 1697.1365 1681.4905 1664.1336 1648.6631 1631.2750 1614.1504 1598.1326
[73] 1591.0377 1587.9573 1587.6974 1576.8235 1562.7157 1562.6217 1562.5767 1551.9625 1528.3684
[82] 1509.1910 1486.9190 1469.3362 1445.6227 1441.4363 1413.5958 1385.7964 1382.8505 1357.9724
[91] 1325.1695 1325.1692 1325.1678 1285.3492 1283.6251 1283.5290 1283.4736 1239.2959 1239.1103
[100] 1239.1102 1193.2691 1189.7019 1189.4728 1189.4653 1147.2802 1143.1372 1110.1379 1110.0581
[109] 1109.9607 1091.7923 1061.2180 1059.8396 1059.8387 1048.7490 1048.7276 1048.6966 1040.8313
[118] 1018.6890 1018.6779 1017.6260 1011.6857 1011.6604 1011.5700 1007.6972 1007.6944 1002.5069
```

RESULTADOS

Resultado 4:
El umbral encontrado a través de chicuadro es de 16.26. Los datos que superan ese valor de dispersión son 5,532; lo cual representa el 1,70%.

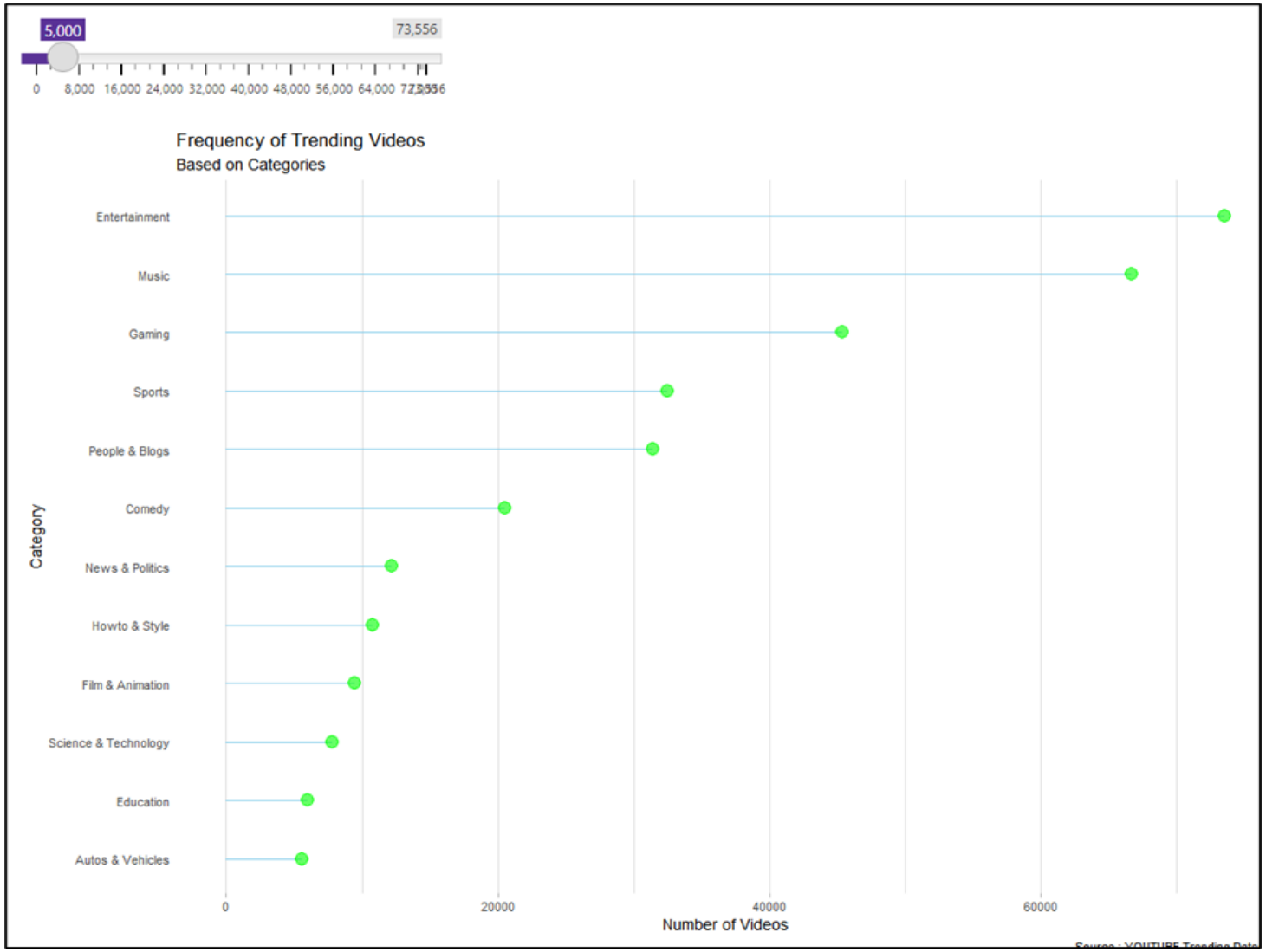


Sin embargo, si la evaluación es mediante la gráfica de cuantil – cuantil, observamos una mayor dispersión en relación al valor de chicuadro . Los resultados se presentan a continuacion.



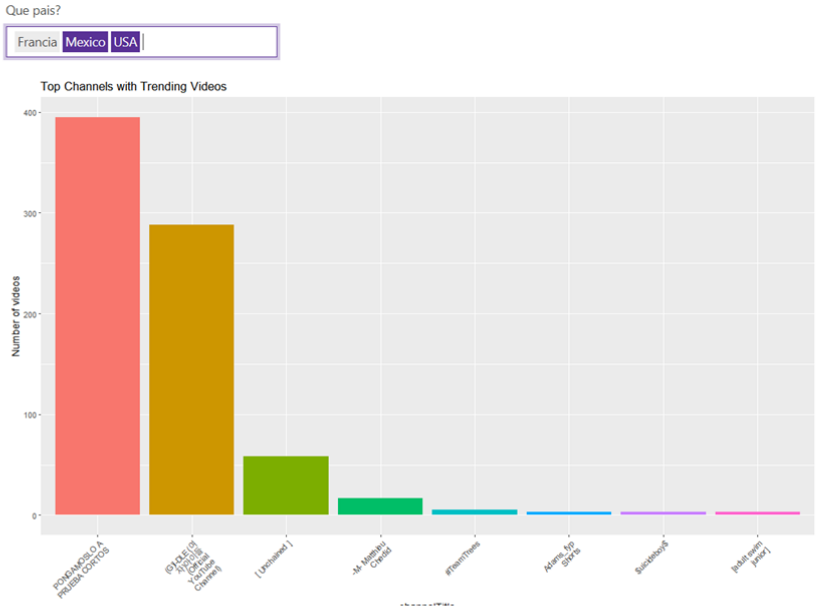
RESULTADOS

Resultado 5: Encontramos en el análisis una mayor inclinación por la generación de videos en las categorías de Entretenimiento y Música, del otro lado del análisis se encuentran Vehículos, Educación, y Ciencia y tecnología.

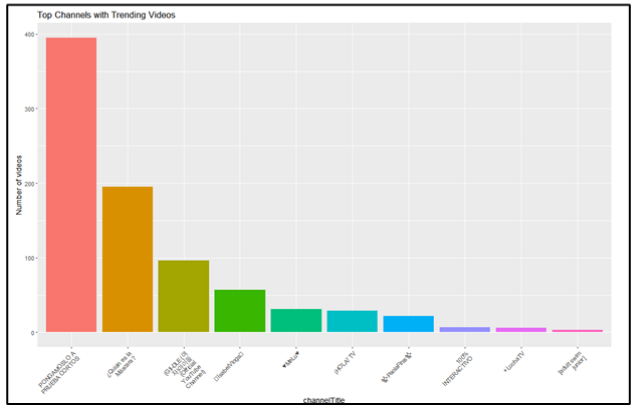


RESULTADOS

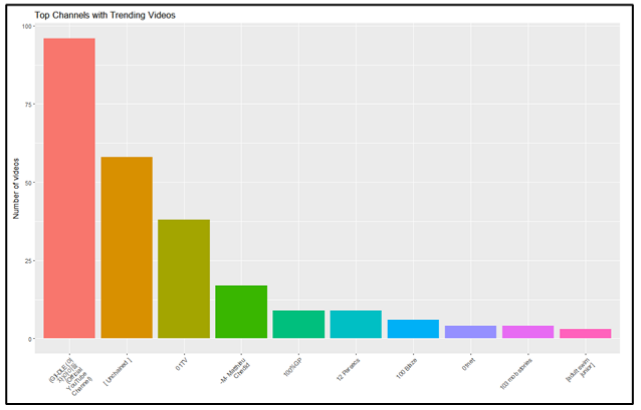
Resultado 6: Es interesante, encontrar que entre los canales que más contenido sube, está en primer lugar “Pongámoslo a prueba cortos”. Lo que sorprende es un canal chino que hace contenido para cada uno de los países, posicionándose como uno de los que más generen en cada uno de ello.



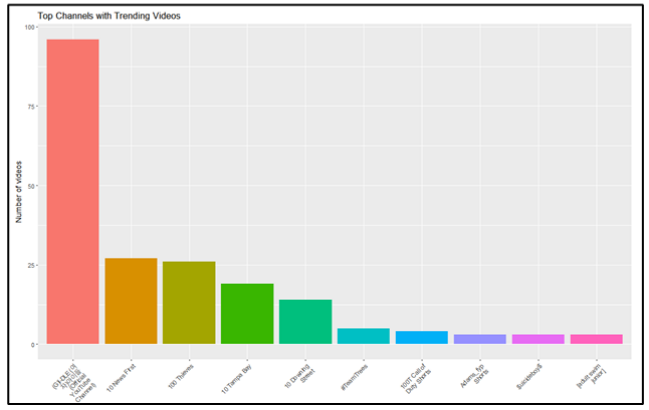
Francia



USA



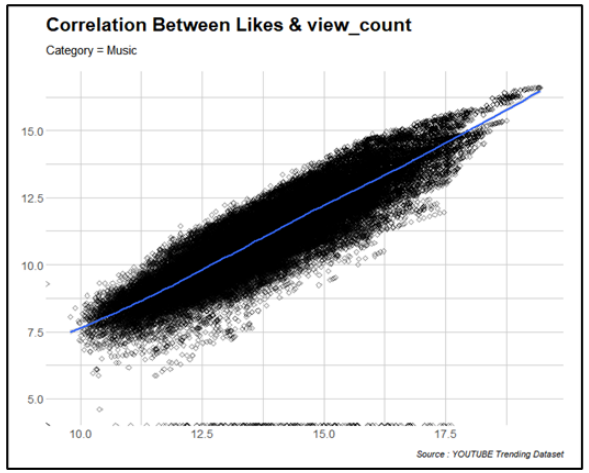
Mexico

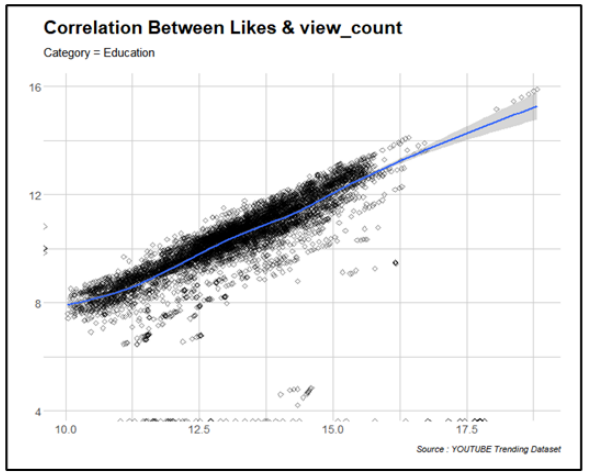


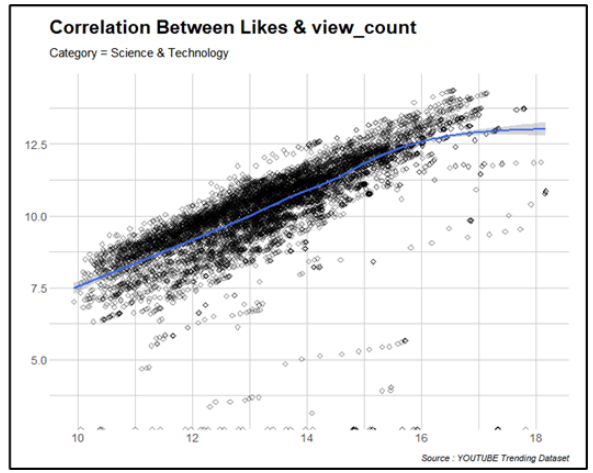
RESULTADOS

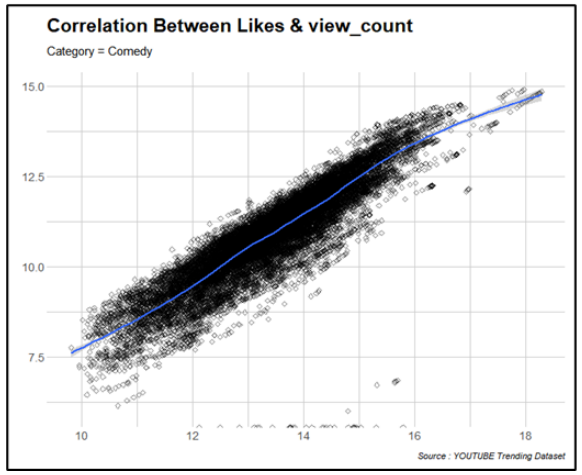
Resultado 7: Por último, hicimos el ejercicio para evaluar la correlación entre el número de reproducciones y LIKES.

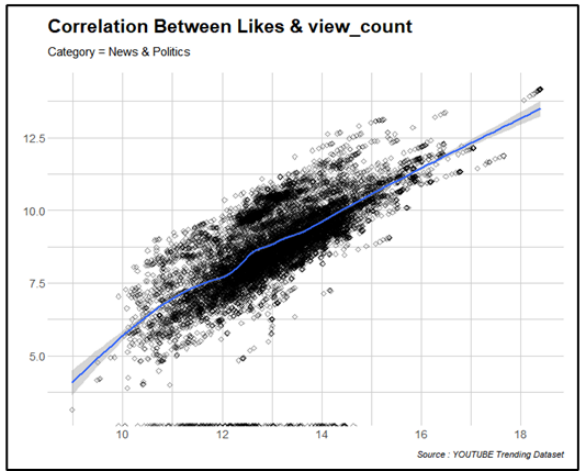
- De primera vista, vemos que las categorías que podríamos llamar divertidas, como “Música” y “Comedy”, tienen una distribución casi homogénea en estas dos variables. No pasa de forma tan notoria en “Educación” y “News Politycs”. Menos aún en “Ciencias y Tecnología”, y “Sin fines de lucro y activismo”.
- Probablemente, estos resultados pueden ayudar a ahondar en la forma psicología del comportamiento de las personas. Que si bien en general, todos tienen una misma dirección en cosas no serias, en temas que son más serios, existe ligera diferencia en la opinión de cada persona

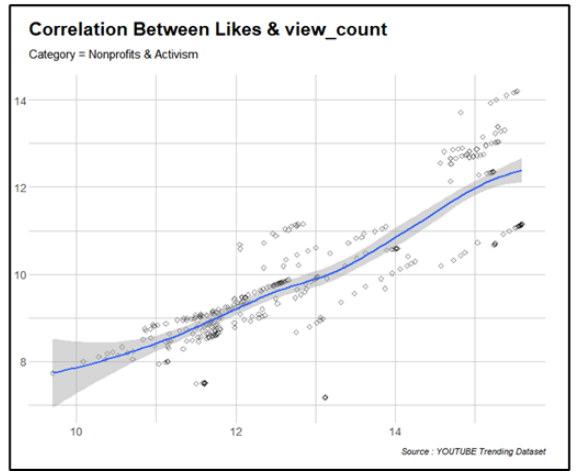












CONCLUSIONES

- La herramienta de MongoDB nos permitió extraer los datos de manera segura y organizada utilizando usuarios de solo lectura con la finalidad de la inmutabilidad de la base de datos.
- Hemos comprobado la facilidad de usar una instancia virtualizada como EC2, la misma que ha ayudado al traslado de altos volúmenes de datos sin interrupción y con la posibilidad de programarse como una tarea repetitiva.
- Hemos constatado la capacidad de R-Studio sobre su capacidad para la conexión a una base de datos como servicio, como es MongoDB, la misma que después de la exploración y procesamiento de la información también ayudo a su presentación de información mediante Shiny Apps.
- Hemos encontrado también una limitación en la conexión a la base de datos como servicio, cuando se trata de la extracción de datos pesados, ya que en oportunidades la carga exigió bastante recurso computacional. Revisar la cantidad de conexiones máximas hacia la BD, es necesario.
- Hemos encontrando que un canal chino genera una cantidad significativa de videos en diferentes países tales como USA, México y Francia. De forma muy diferente al comportamiento de otros canales que tienen mayor presencia en ciertos países.
- Menos del 0.01% podemos decir que son videos que han logrado una variación representativa al resto. Si lo vemos en números, existe una probabilidad de 0.01%. de que si generas un video alcance un gran éxito.

BIBLIOGRAFÍA

A continuación, mostramos los recursos que nos permitieron completar este trabajo

- A. N. PETTITT, A two-sample Anderson-Darling rank statistic, Biometrika, Volume 63, Issue 1, 1976, Pages 161–168, https://doi.org/10.1093/biomet/63.1.161
- Emad-Eldin, A. A., & Öztürk, A. (1988). A modified one-sample QQ plot and a test for normality. Journal of Statistical Computation and Simulation, 29(1), 1-15.
- G. Fasano, A. Franceschini, A multidimensional version of the Kolmogorov–Smirnov test, Monthly Notices of the Royal Astronomical Society, Volume 225, Issue 1, March 1987, Pages 155–170, https://doi.org/10.1093/mnras/225.1.155
- Hanusz, Z., & Tarasińska, J. (2015). Normalization of the Kolmogorov–Smirnov and Shapiro–Wilk tests of normality. Biometrical Letters, 52(2), 85-93.
- Lee, R., Qian, M., & Shao, Y. (2014). On rotational robustness of Shapiro-Wilk type tests for multivariate normality. Open Journal of Statistics, 4(11), 964.
- Peter A. W. Lewis. (1961). Distribution of the Anderson-Darling Statistic. The Annals of Mathematical Statistics, 32(4), 1118–1124. http://www.jstor.org/stable/2237910
- Ronald L. Iman (1982) Graphs for use with the Lilliefors Test for Normal and Exponential Distributions, The American Statistician, 36:2, 109-112, DOI: 10.1080/00031305.1982.10482799
- Walfish, S. (2006). A review of statistical outlier methods. Pharmaceutical technology, 30(11), 82.

Gracias