Universidad Peruana de Ciencias Aplicadas

Trabajo Final

Curso: Data Science

Sección: CC52

Profesor: Nérida Isabel Manrique Tunque

Código del alumno	**Apellidos y Nombres completos**
U202115844	Mancilla Cienfuegos, Paula Jimena
U202110078	Loyola Huaman, Jose Alejandro
U20201E766	Vargas Soto, Lennin Jhair
U20231D424	Caro Leon, Lesly Estefany

Noviembre, 2023

I. INTRODUCCIÓN

Hoy en día el mundo de la analítica de datos está en constante evolución además es una herramienta de una importancia significativa para las decisiones estratégicas en un negocio. Teniendo en cuenta la relevancia del manejo de los datos, nuestro proyecto nace como respuesta a la solicitud de una consultora internacional con sede en Lima que nos propone desarrollar un análisis de un conjunto de datos con la finalidad de conocer las tendencias de los videos de Youtube en Gran Bretaña. El objetivo principal de este proyecto es obtener una comprensión en profundidad de los patrones de visualización, las preferencias de contenido y otros indicadores clave de desempeño que caracterizan el comportamiento de la audiencia en línea.

II. INTEGRANTES DEL GRUPO Y ROLES

Para asignar los roles a cada integrante de mi grupo organizamos una reunión en Meet donde cada uno eligió que parte de la Metodología Crisp-DM se encargaría. A continuación se presenta un cuadro donde se mostrará los roles de cada integrante

Nombre	Rol	Tareas Asignadas
Alejandro Loyola	Business Project Sponsor	Comprensión y descripción de los objetivos
Lesly Caro	Data Engineer	En esta fase se busca comprender la estructura y las características clave del conjunto de datos, explorar patrones y relaciones en los datos, y evaluar su calidad para su uso en análisis más detallados o modelado.
Jhair	Data Engineer	En esta fase se ejecuta la preparación de los datos para adaptarlos a las técnicas/algoritmos de Data Mining que serán seleccionadas.
Paula Mancilla	Data Analytic	En esta fase se busca establecer las técnicas de modelado más apropiadas para el proyecto de Data Mining específico de acuerdo con los objetivos planteados.

III. METODOLOGÍA CRISP-DM

1. COMPRENSIÓN DEL NEGOCIO

Los objetivos del proyecto son los siguientes:

1. Comprender las tendencias de visualización
2. Facilitar la toma de decisiones
3. Mejorar la retención de la audiencia

OBJETIVOS DEL PROYECTO

Identificar y comprender las preferencias de visualización de videos en la plataforma YouTube en siete países específicos, con el propósito de proporcionar información valiosa que respalde las estrategias de marketing digital, la creación de contenido y la toma de decisiones empresariales, permitiendo una mejor personalización de las estrategias y el aprovechamiento de oportunidades de mercado en cada uno de estos países.

OBJETIVOS DE DATA SCIENCE

Los objetivos del negocio son los siguientes:

- Encontrar qué categorías de videos son los de mayor tendencia
- Mostrar el cambio de volumen de los videos en tendencia a lo largo del tiempo
- Encontrar la frecuencia de búsqueda de los canales de Youtube
- Hallar los estados que presentan el mayor número de vistas, “me gusta” y “no me gusta”
- Desarrollar un modelo predictivo mediante técnicas de data mining que permita predecir si un video se convertirá en tendencia en YouTube.

2. COMPRENSIÓN DE LOS DATOS

Descripción General de los Datos

- **Cantidad de Observaciones (Filas):** 38,916
- **Cantidad de Características (Columnas):** 20
- **Columnas con Más Frecuencia de Aparición:**
    - **video_id**: El video con ID **Il-an3K9pjg** aparece 38 veces.
    - **channel_title**: El canal "The Tonight Show Starring Jimmy Fallon" aparece 208 veces.
    - Otros valores más comunes incluyen fechas de tendencia, etiquetas y horas de publicación.
- **Estadísticas Descriptivas (para datos numéricos):**
    - Vistas (**views**), likes (**likes**), dislikes (**dislikes**), etc., tienen un rango amplio, desde valores muy bajos hasta muy altos.

Tipo de Datos

Número	Nombre de Columna	Tipo de Dato	Descripción de Columna
1	video_id	object	Identificador único del video.
2	trending_date	object	Fecha en la que el video se volvió tendencia.
3	title	object	Título del video.
4	channel_title	object	Nombre del canal que subió el video.
5	category_id	int64	ID de la categoría del video.
6	publish_time	object	Fecha y hora de publicación del video.
7	tags	object	Etiquetas asociadas al video.

8	views	int64	Número de visualizaciones del video.
9	likes	int64	Cantidad de 'me gusta' del video.
10	dislikes	int64	Cantidad de 'no me gusta' del video.
11	comment_count	int64	Número de comentarios en el video.
12	thumbnail_link	object	Enlace a la miniatura del video.
13	comments_disabled	bool	Si los comentarios están deshabilitados.
14	ratings_disabled	bool	Si las valoraciones están deshabilitadas.
15	video_error_or_removed	bool	Si el video fue eliminado o tuvo errores.
16	description	object	Descripción del video.
17	state	object	Estado de ubicación relacionado con el video.
18	lat	float64	Latitud geográfica asociada con el estado.
19	lon	float64	Longitud geográfica asociada con el estado.

20	geometry	object	Datos de geometría (generalmente coordenadas).

- Datos numéricos (como **views**, **likes**, **dislikes**).
- Datos categóricos/objetos (como **video_id**, **title**, **channel_title**).
- Datos booleanos (como **comments_disabled**, **ratings_disabled**).

Valores Nulos

- **Presencia de Valores Nulos:**
  - La única columna con valores nulos es **description**, con 612 valores nulos.

Distribución de Vistas, Likes, Dislikes y Comentarios: Histogramas para visualizar cómo se distribuyen estas métricas entre los videos.

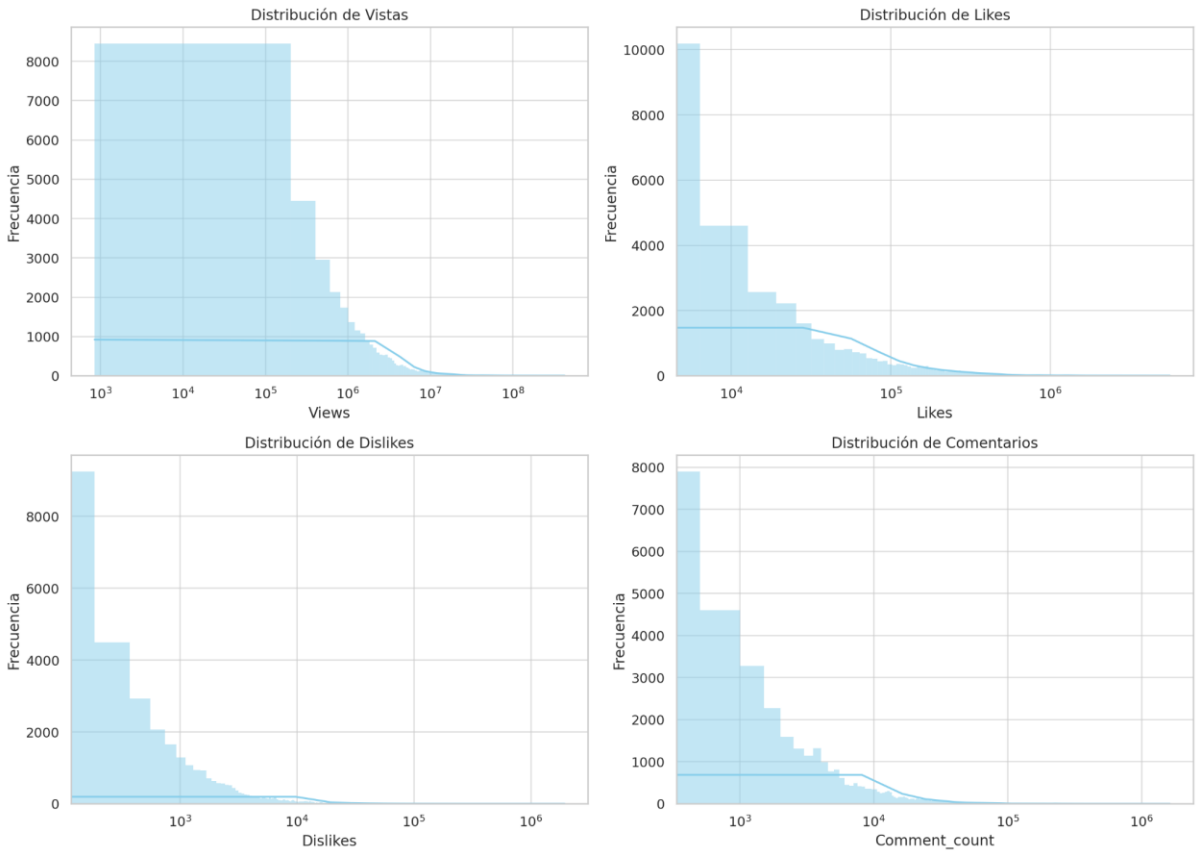


Imagen 1: Gráfico de distribución de Vistas, Likes, Dislikes y Comentarios **(Elaboración Propia)**

Interpretación:

- **Distribución de Vistas:** La mayoría de los videos tienen un número relativamente bajo de vistas, con solo unos pocos alcanzando cifras muy altas.
- **Distribución de Likes y Dislikes:** Un patrón similar se observa con los likes y dislikes, donde la mayoría de los videos tienen una cantidad menor de reacciones.
- **Distribución de Comentarios:** La cantidad de comentarios también sigue un patrón similar, con la mayoría de los videos recibiendo relativamente pocos comentarios.

Relación entre Vistas y Likes/Dislikes: Gráfico de dispersión para entender la relación entre el número de vistas de un video y la cantidad de likes o dislikes que recibe.

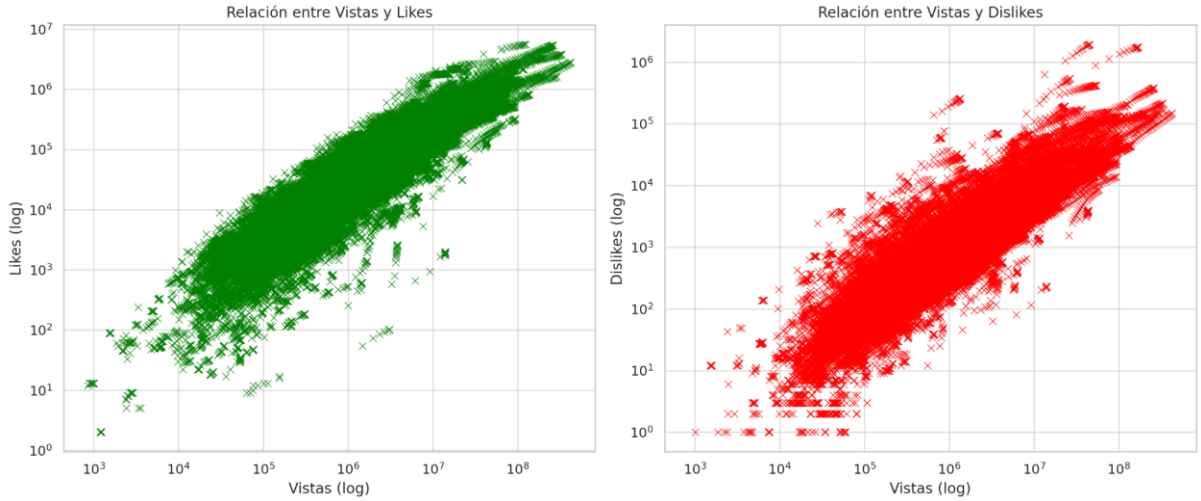


Imagen 2: Gráfico de relación entre Vistas y Likes/Dislikes **(Elaboración Propia)**

Interpretación:

· Relación entre Vistas y Likes: Se observa una tendencia positiva en la que, a medida que aumentan las vistas, también lo hacen los 'me gusta'. Este patrón sugiere una correlación entre la popularidad de un video (medida en vistas) y la cantidad de reacciones positivas (likes) que recibe. Es lógico que los videos más vistos también tengan más oportunidades de recibir 'me gusta'. La relación no es perfectamente lineal, lo que indica que no todos los videos populares reciben likes en la misma proporción.

· Relación entre Vistas y Dislikes: Similar a la relación entre vistas y likes, hay una correlación positiva entre las vistas y los 'no me gusta'. Aunque los dislikes son generalmente menos frecuentes que los likes, esta tendencia también indica que los videos más vistos tienden a recibir más 'no me gusta'. Esto puede deberse a que tienen una audiencia más amplia y, por lo tanto, más exposición a opiniones variadas. Sin embargo, la dispersión en este gráfico es un poco más amplia, lo que sugiere que los factores que conducen a un 'no me gusta' pueden ser más variados o menos directamente relacionados con el número de vistas que en el caso de los 'me gusta'.

Calidad de los Datos

1. **Tipos de Datos:**
    - Los tipos de datos son consistentes con lo esperado. Por ejemplo, **views**, **likes**, **dislikes**, y **comment_count** son numéricos, mientras que **video_id**, **title**, etc., son objetos (cadenas de texto).
2. **Valores Nulos:**
    - La única columna con valores nulos significativos es **description**, con 612 valores nulos, lo que representa aproximadamente el 1.57% del total de filas.
    - Todos los demás campos no tienen valores nulos o su proporción es insignificante.
3. **Outliers en 'views':**
    - Se identificaron 5,308 valores atípicos en la columna **views**, basados en el método del Rango Intercuartílico (IQR). Estos pueden representar videos extremadamente populares o con muy pocas vistas.

Interpretación

- **Valores Nulos:** La presencia de valores nulos en **description** puede ser normal, ya que algunos videos pueden no tener una descripción. Sin embargo, es importante considerar esto al realizar análisis que involucren esta columna.
- **Outliers:** La detección de valores atípicos en **views** es importante, ya que estos valores pueden distorsionar el análisis. Dependiendo del objetivo del análisis, puedes optar por excluir estos outliers, analizarlos por separado, o tomar en cuenta su impacto en el análisis general.

3. PREPARACIÓN DE LOS DATOS

Primero a la hora de preprocesar los datos identificamos los valores faltantes en las columnas que para nuestro caso tenemos 612 datos faltantes en “description”:

```
#Mostrar los datos faltantes
missing_values = df.isnull().sum()
print("\nValores faltantes en cada columna:")
print(missing_values)
```

```
Valores faltantes en cada columna:
video_id                    0
trending_date               0
title                       0
channel_title               0
category_id                 0
publish_time                0
tags                        0
views                       0
likes                       0
dislikes                    0
comment_count               0
thumbnail_link              0
comments_disabled           0
ratings_disabled            0
video_error_or_removed      0
description               612
state                       0
lat                         0
lon                         0
geometry                    0
dtype: int64
```

Imagen 3: Visualización de datos faltantes **(Elaboración Propia)**

Utilizamos la técnica de eliminación, ya que al ser una descripción de un video de youtube consideramos que es la mejor opción.

```
# Eliminar filas con valores faltantes en 'description'
df_clean = df.dropna(subset=['description'])
```

Imagen 4: Eliminación de valores faltantes en “Description” **(Elaboración Propia)**

Luego hacemos un boxplot para identificar los outliers en las variables numéricas, en el gráfico se ve que en los views hay valores muy diferentes, pero al ser una video de Youtube este mismo puede tener diferentes números de vistas.

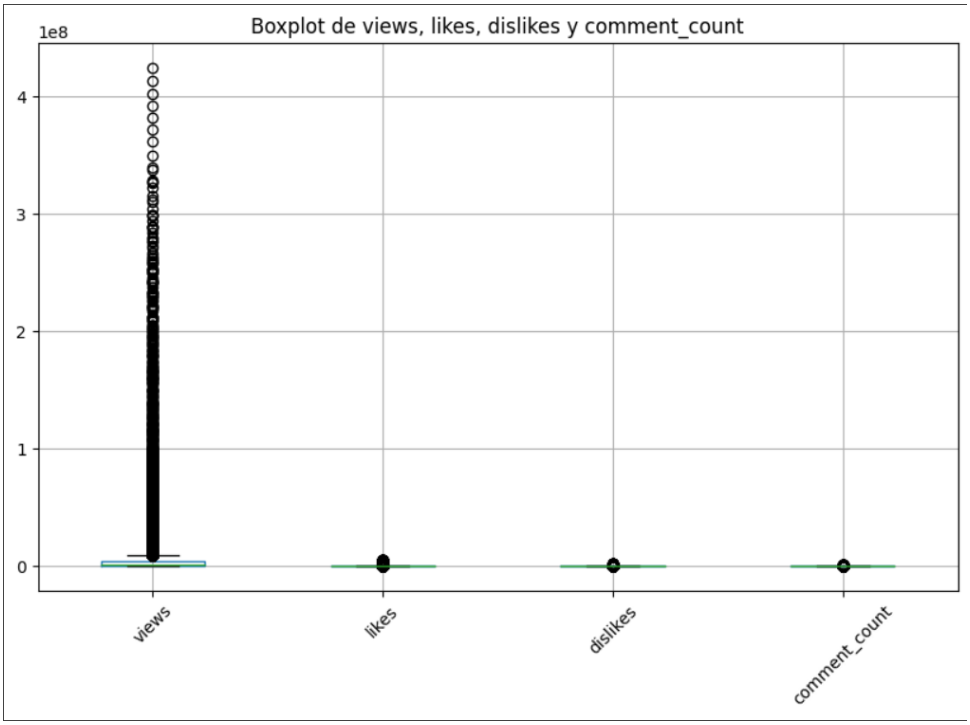


Imagen 5: Boxplot de valores numéricos **(Elaboración Propia)**

Así que identificamos si hay valores negativos por un error de digitación o algún otro error, ya que el número de vistas debe ser positivo.

```python
# Seleccionar las columnas relevantes para la verificación de números negativos
columnas_relevantes = ['views', 'likes', 'dislikes', 'comment_count']

# Verificar si hay números negativos en las columnas seleccionadas
for columna in columnas_relevantes:
    negativos = df_clean[df_clean[columna] < 0]
    cantidad_negativos = len(negativos)
    if cantidad_negativos > 0:
        print(f"Se encontraron {cantidad_negativos} números negativos en la columna {columna}.")
    else:
        print(f"No se encontraron números negativos en la columna {columna}.")
```

```
No se encontraron números negativos en la columna views.
No se encontraron números negativos en la columna likes.
No se encontraron números negativos en la columna dislikes.
No se encontraron números negativos en la columna comment_count.
```

Imagen 6: Verificación de valores negativos **(Elaboración Propia)**

A continuación, convertimos la variable “trending_date” a un formato más fácil y conocido para trabajar.

```
# Convertir la columna 'trending_date' a formato de fecha
df_clean['trending_date'] = pd.to_datetime(df_clean['trending_date'], format='%y.%d.%m')
print(df_clean.head())
```

Imagen 7: Transformación de la columna “trending_date” **(Elaboración Propia)**

Por último, creamos una nueva variable llamada “category_name” la cual es creada a partir de la variable “category_id” del dataset y la variable “tittle” del archivo json.

```
import json

with open('/content/GB_category_id.json') as f:
    data = json.load(f)

category_dict_json = {int(item['id']): item['snippet']['title'] for item in data['items']}

df_clean['category_name'] = df_clean['category_id'].map(category_dict_json)

print(df_clean[['category_id', 'category_name']].head(10))
```

```
   category_id    category_name
0           26    Howto & Style
1           24    Entertainment
2           10            Music
3           17           Sports
4           25  News & Politics
5           24    Entertainment
6           10            Music
7           22   People & Blogs
8           10            Music
9           10            Music
<ipython-input-121-6d0716168591>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_clean['category_name'] = df_clean['category_id'].map(category_dict_json)
```

Imagen 8: Creación de la columna “category_name” **(Elaboración Propia)**

Para responder a las siguientes preguntas se creó una visualización para cada una de ellas:

1. ¿Qué categorías de videos son las de mayor tendencia?

Como se observa en el gráfico de barras las 2 categorías de mayor tendencia son Music y Entertainment.

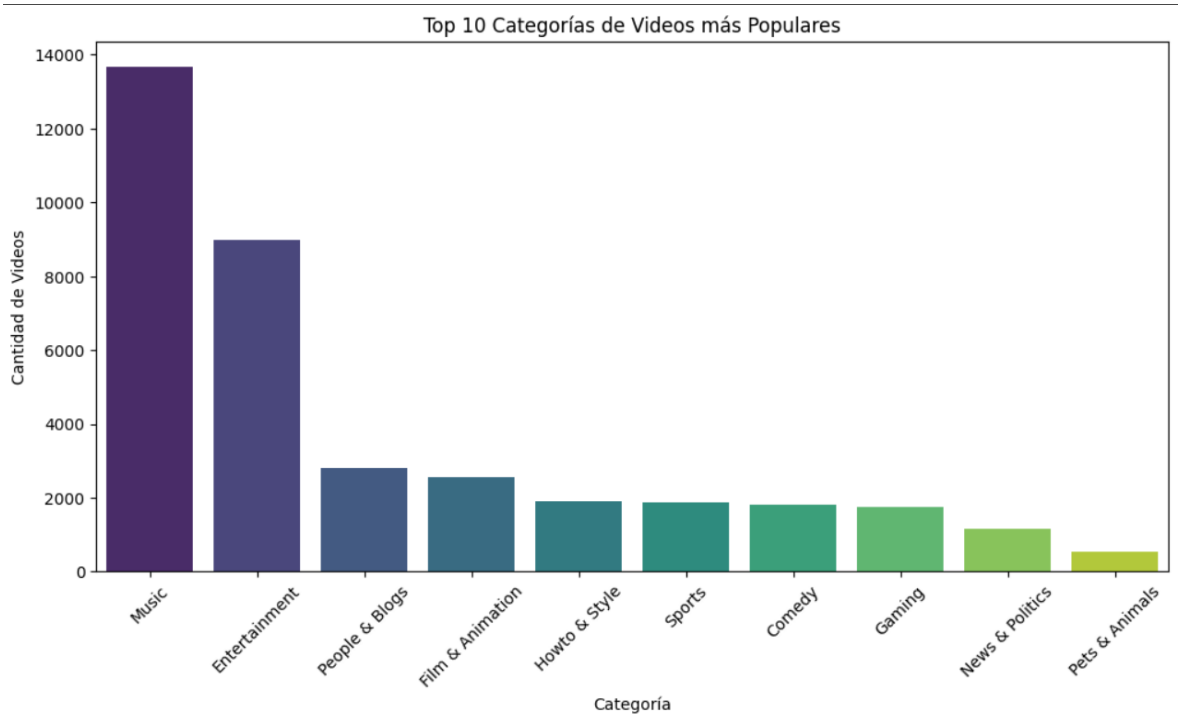


Imagen 9: Gráfico de barras de categorías de vídeos más populares **(Elaboración Propia)**

2. ¿Qué categorías de videos son los que más gustan? ¿Y las que menos gustan?

Como se observa en este primer gráfico las categorías que más gustan por la cantidad de likes que tienen son Music y Entertainment.

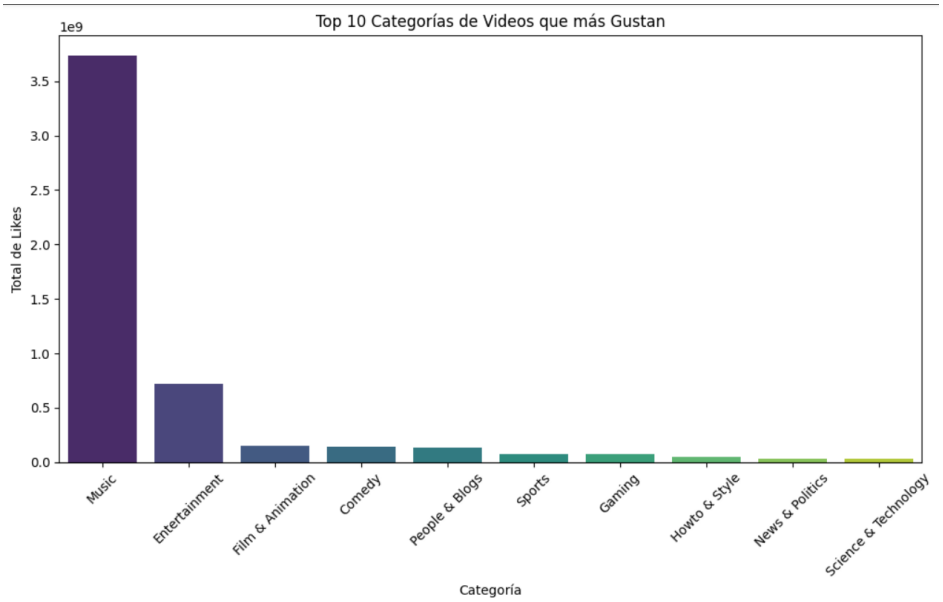


Imagen 10: Gráfico de barras de categorías de vídeos que más gustan **(Elaboración Propia)**

Y como se observa en el segundo gráfico las categorías que menos gustan son Shows y Travel & Events.

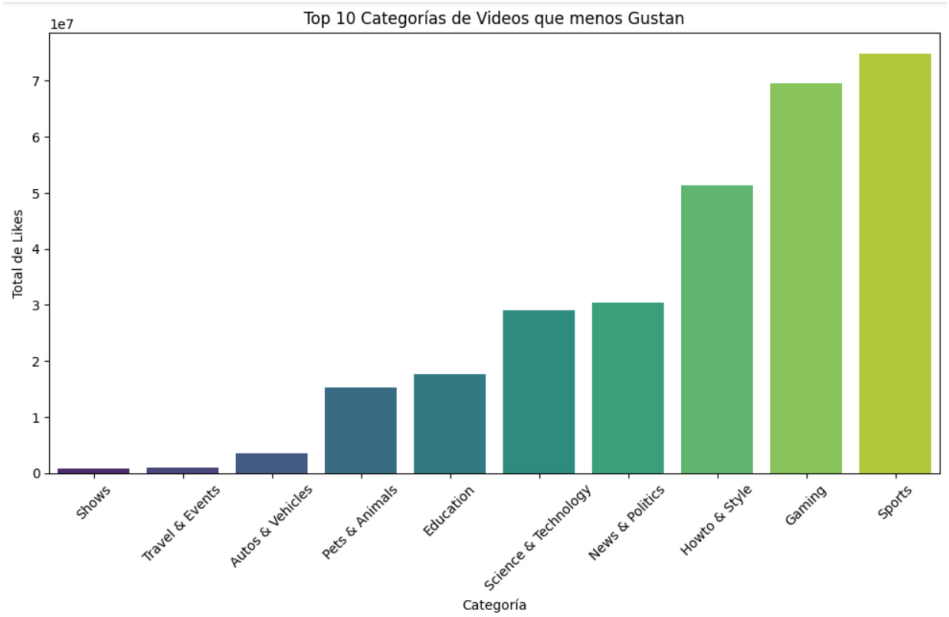


Imagen 11: Gráfico de barras de categorías de vídeos que menos gustan **(Elaboración Propia)**

3. ¿Qué categorías de videos tienen la mejor proporción (ratio) de “Me gusta” / “No me gusta”?

En el gráfico se muestra que las categorías de Shows y Pets & Animals son las que mejor ratio “Me gusta” / “No me gusta” tienen.

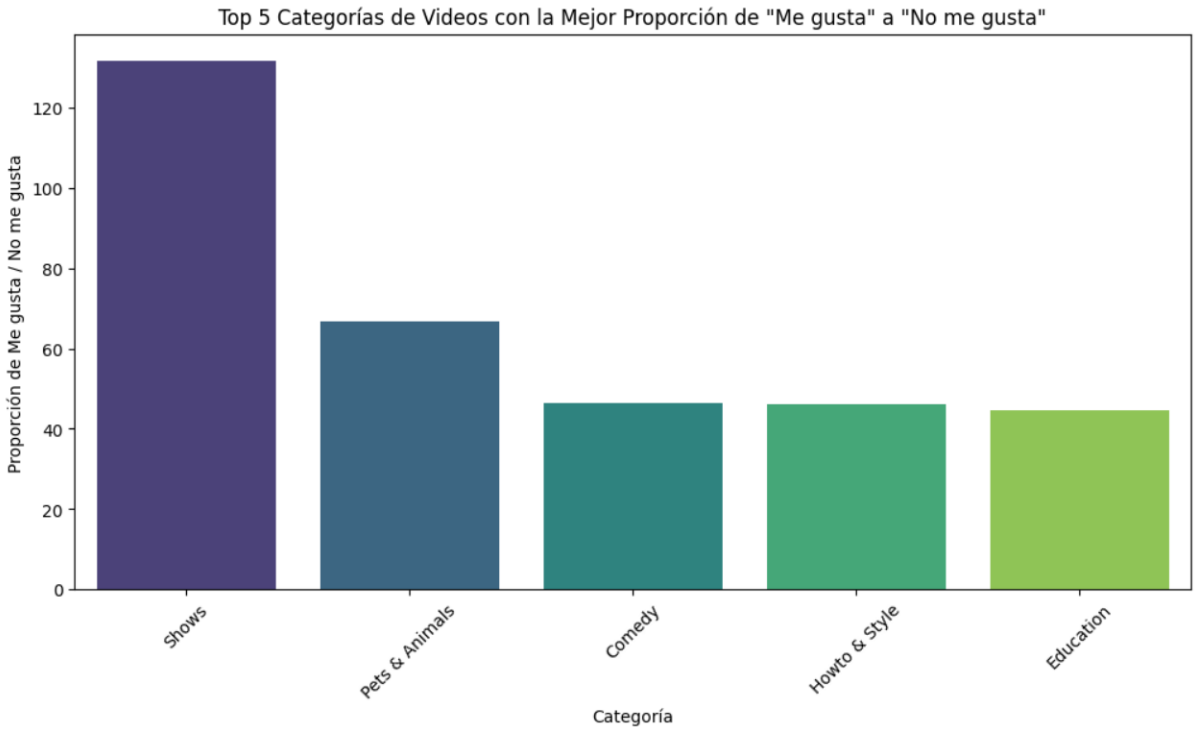


Imagen 12: Gráfico de categorías con mejor ratio “me gusta”/”no me gusta” **(Elaboración Propia)**

4. ¿Qué categorías de videos tienen la mejor proporción (ratio) de “Vistas” / “Comentarios”?

En el gráfico se muestra que las categorías de Travel & Events y Comedy son las que mejor ratio “Vistas” / “Comentarios” tienen.

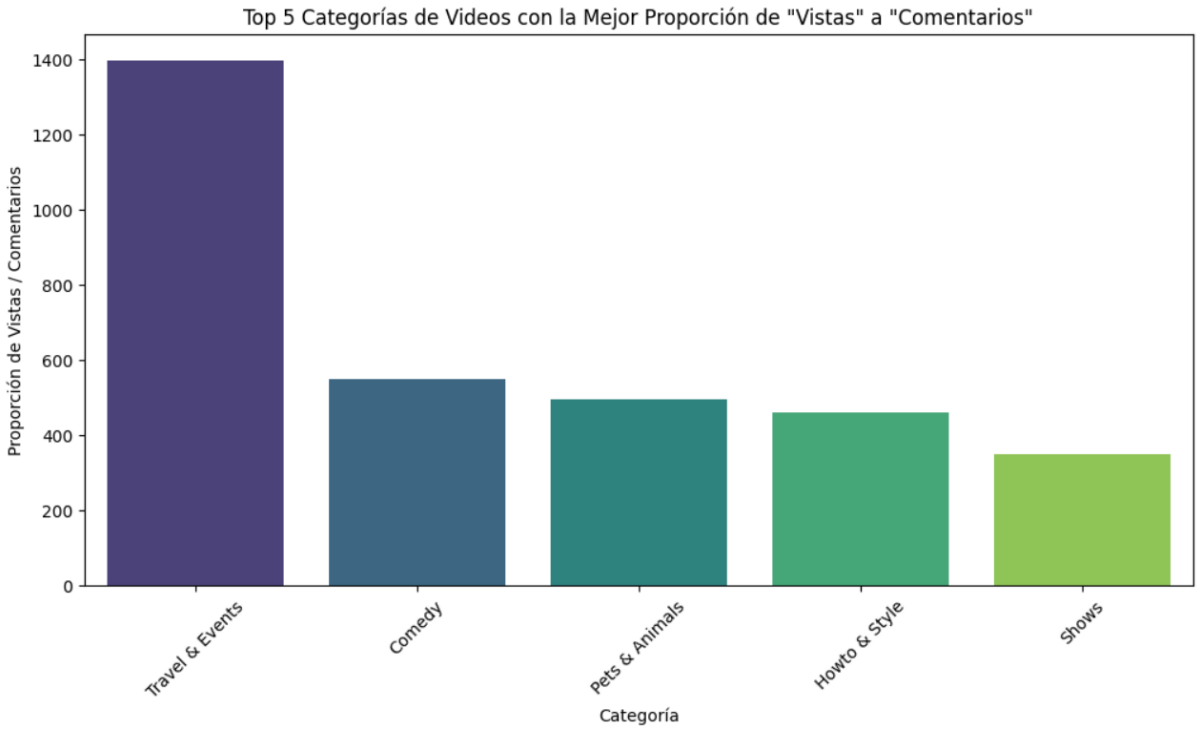


Imagen 13: Gráfico de categorías con mejor ratio “vistas”/”comentarios” **(Elaboración Propia)**

5. ¿Cómo ha cambiado el volumen de los videos en tendencia a lo largo del tiempo?

En el gráfico se muestra que desde el 2017-12 hasta el 2018-03 se ha estado manteniendo el volumen de videos en tendencia. Luego de esto en 2018-04 en adelante se ha visto varios cambios bruscos y el pico más bajo.

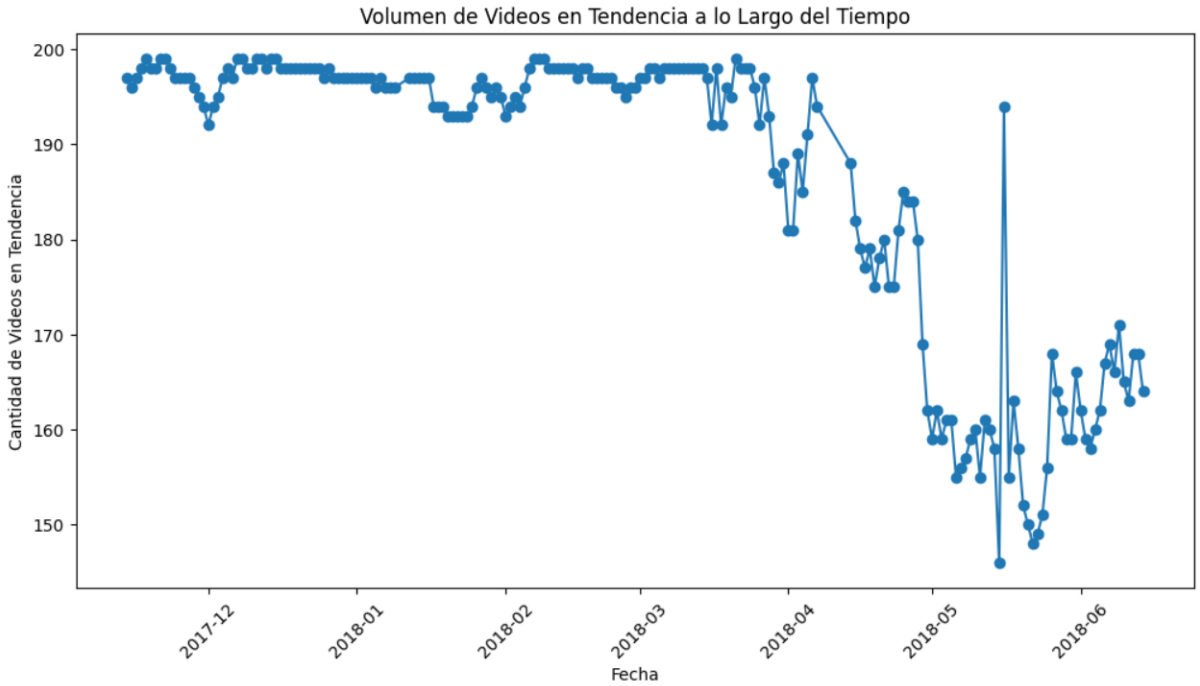


Imagen 14: Gráfico de volumen de vides de tendencia a lo largo del timepo **(Elaboración Propia)**

6. ¿Qué canales de YouTube son tendencia más frecuentemente? ¿Y cuáles con menos frecuencia?

En el primer gráfico se observa que los canales de Youtube más frecuentes en la Lista de Videos de tendencia son: “the tonight show with jimmy fallon”, “TheEllenShow” y “Jimmy Kimmel Live”.

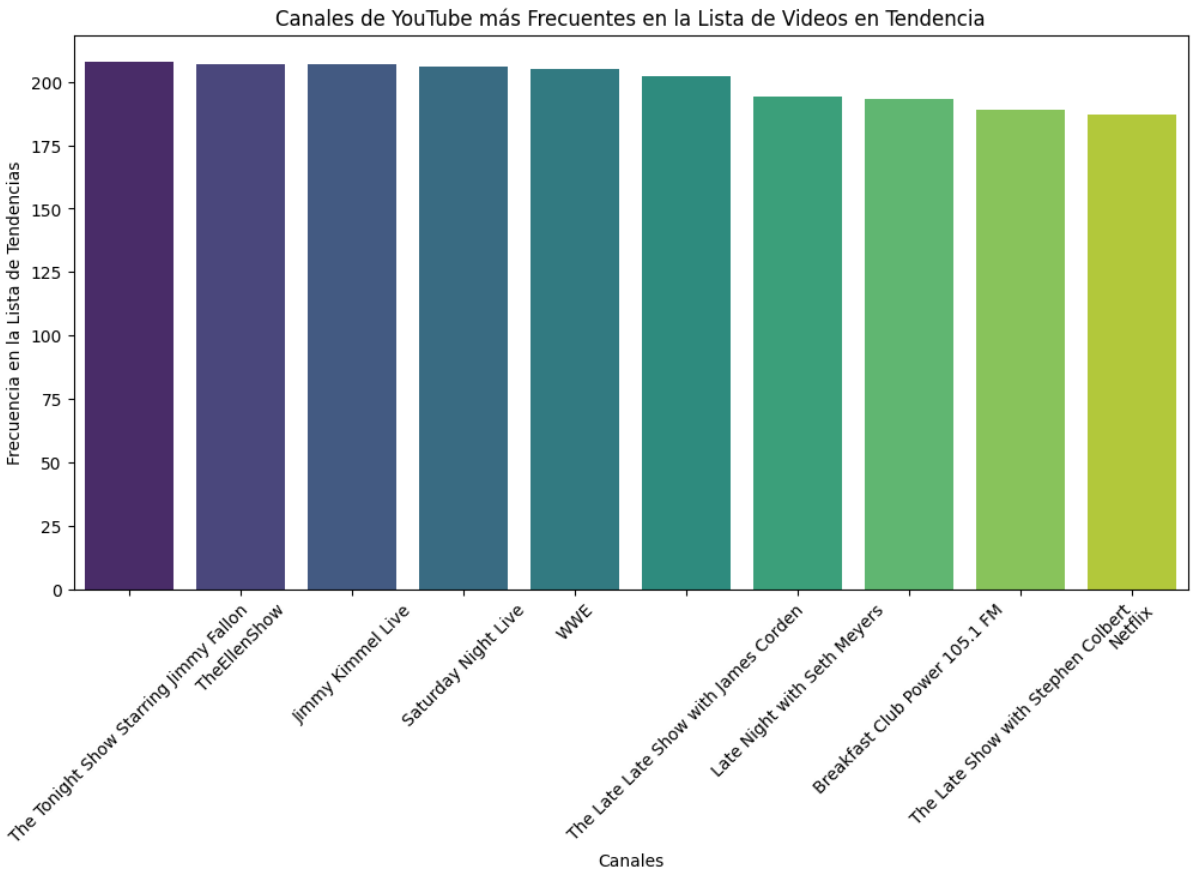


Imagen 15: Gráfico de canales de youtube más frecuentes en la lista de videos en tendencia **(Elaboración Propia)**

En el segundo gráfico se observa que los canales de Youtube menos frecuentes en la Lista de Videos de tendencia son: “joanjetsetter”, “FPFutebolOficial”, “The Chainsmokers”, entre otros más.

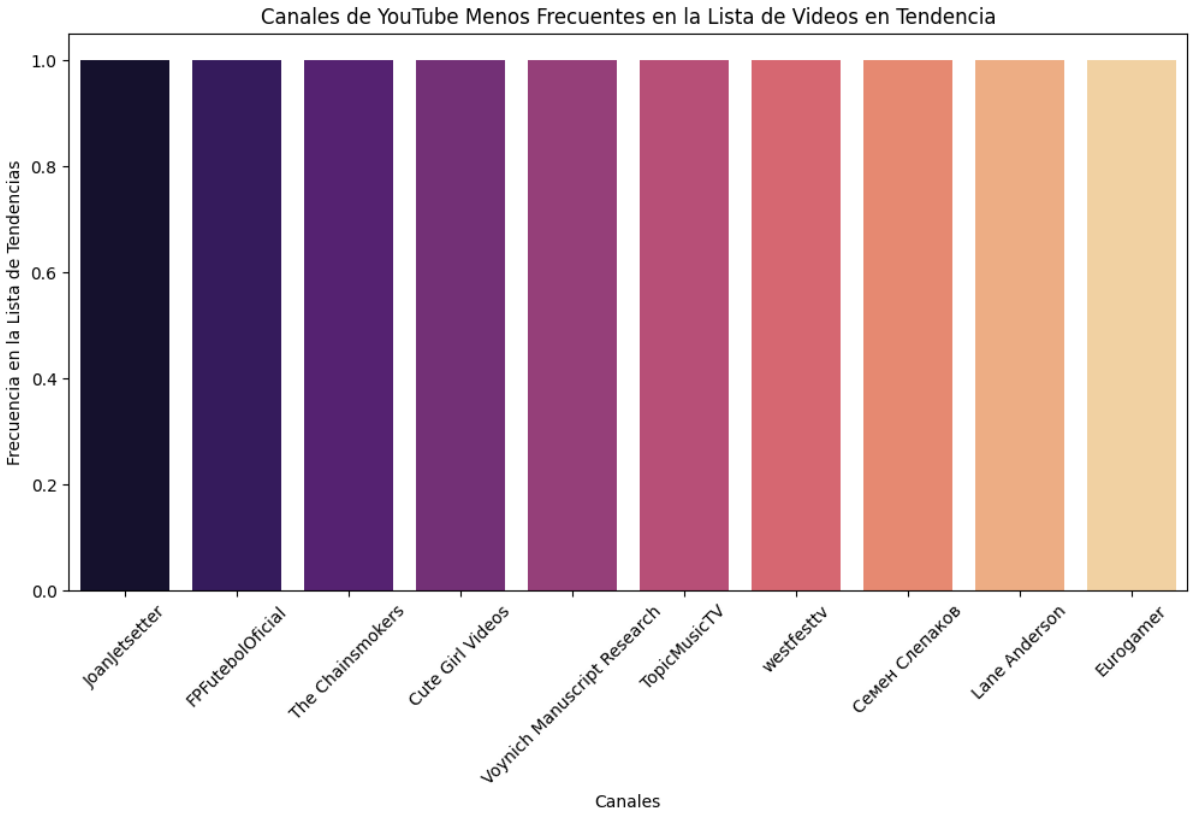


Imagen 16: Gráfico de canales de youtube menos frecuentes en la lista de videos en tendencia **(Elaboración Propia)**

7. ¿En qué Estados se presenta el mayor número de “Vistas”, “Me gusta” y “No me gusta”?

Como se observa en el primer gráfico los 5 estados con más vistas en Gran Bretaña son: “Southampton”, “Aberdeen”, “Stockton On Tees”, “South Yorkshire” y “West Midlands”.

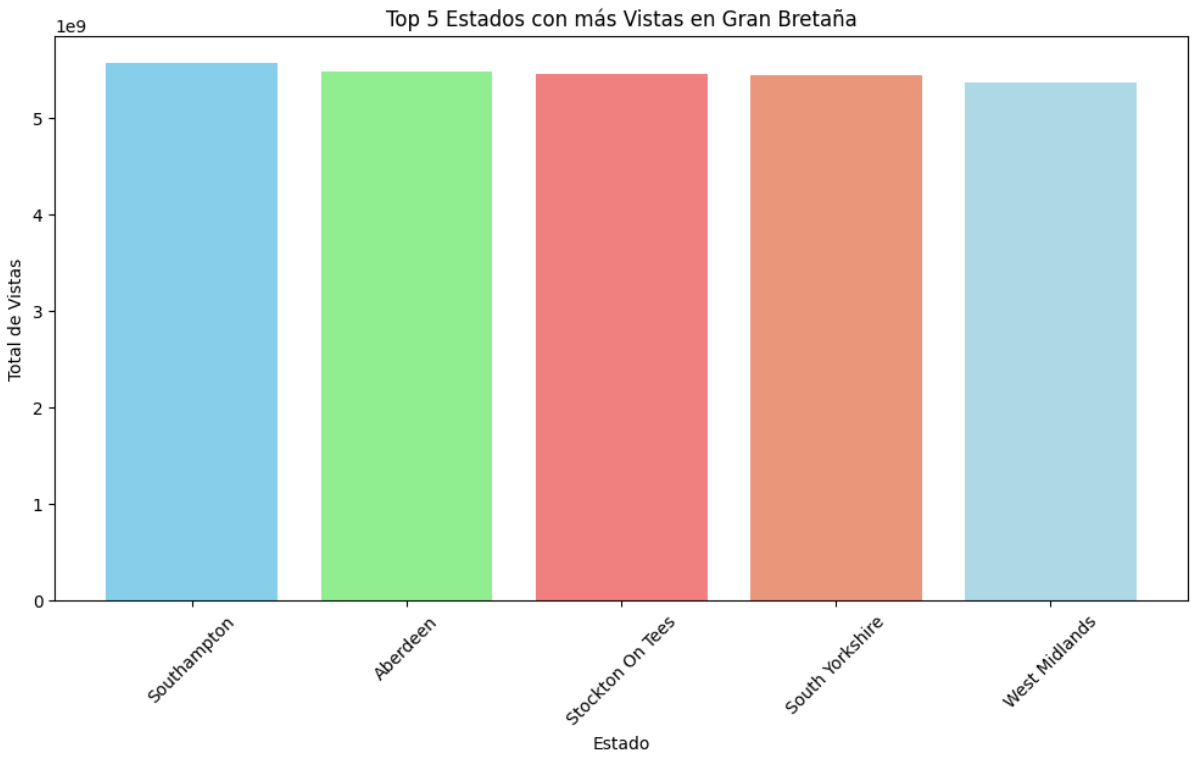


Imagen 17: Gráfico de estados con más vistas en Gran Bretaña **(Elaboración Propia)**

Como se observa en el segundo gráfico los 5 estados con más me gusta en Gran Bretaña son: “South Yorkshire”, “West Midlands”, “Leicester”, “Devon” y Perthshire And Kinross”.

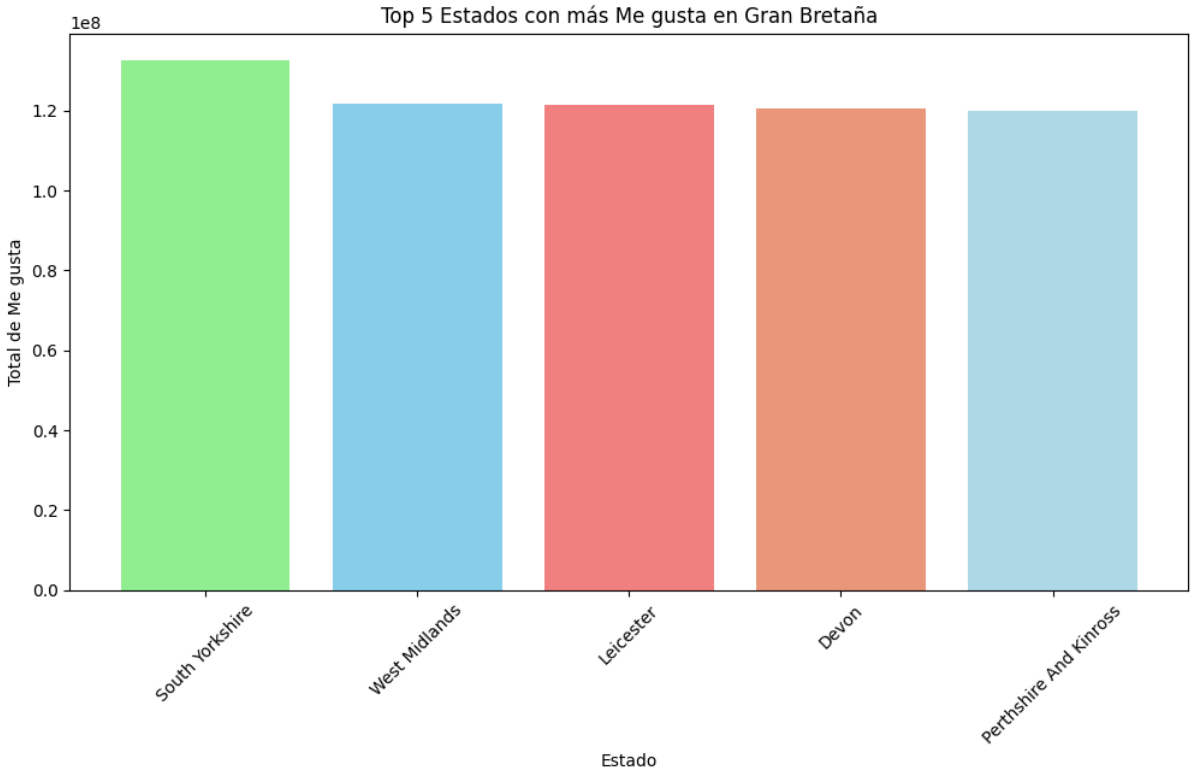


Imagen 18: Gráfico de estados con más me gusta en Gran Bretaña **(Elaboración Propia)**

Como se observa en el tercer gráfico los 5 estados con más no me gusta en Gran Bretaña son: “Bath and North East Somerset ”, “Southampton”, “South Yorkshire”, “Inverclyde” y “Glasgow”.

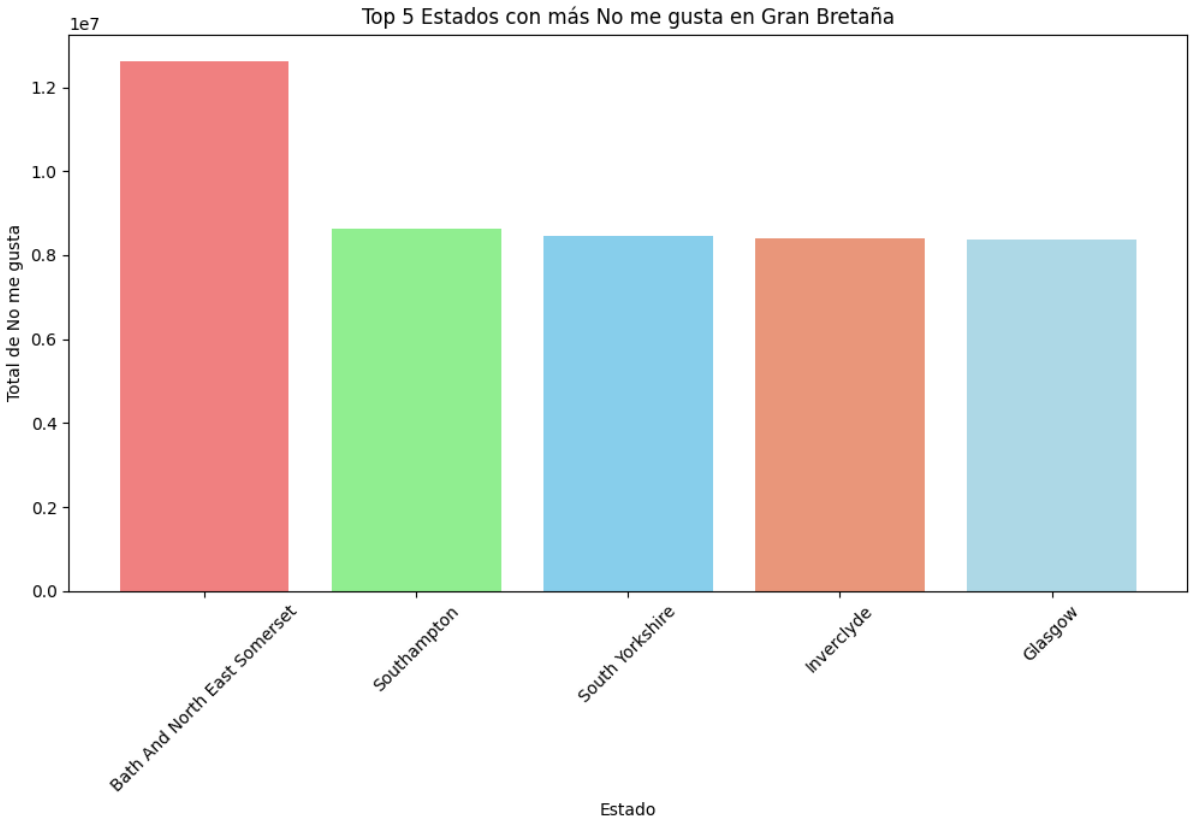


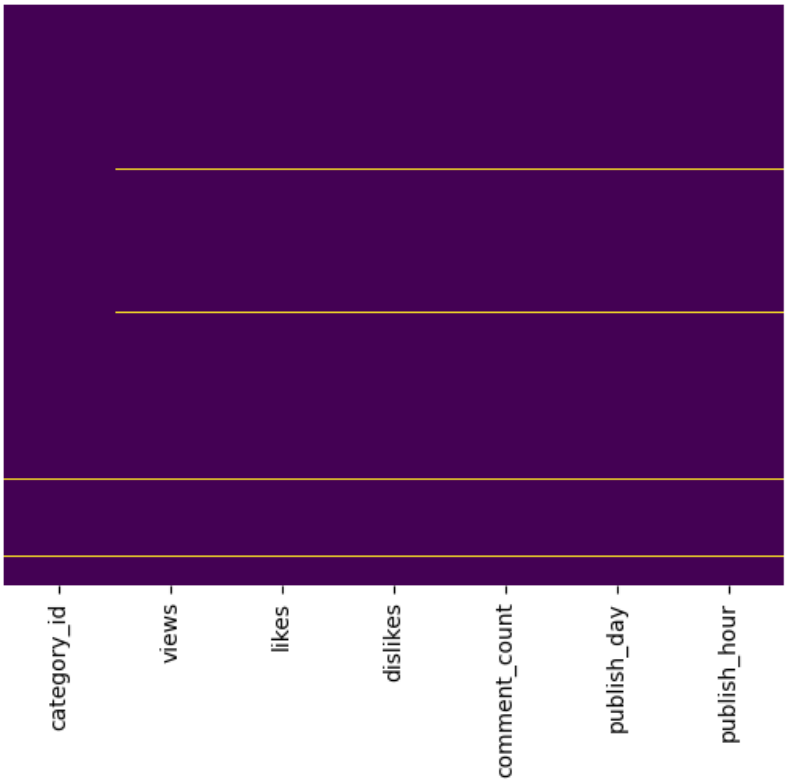
Imagen 19: Gráfico de estados con más no me gusta en Gran Bretaña **(Elaboración Propia)**

4. MODELIZAR Y EVALUAR LOS RESULTADOS-> DATA SCIENCE

1. **Definir Objetivos y Preguntas de Investigación**
   Se utiliza cuando la variable de respuesta es binaria (dos clases) o categórica ordinal. En este caso la variable categórica fue category
   Es adecuada para problemas de clasificación, donde el objetivo es asignar una observación a una categoría. Como el caso es de predecir algo binario, entonces usar Regresión logística es la mejor opción considerar la regresión logística.
2. **Análisis Exploratorio de Datos (EDA)**
   Realizamos un análisis exploratorio de datos detallado para comprender la distribución de las variables. Es así como seleccionamos las variables 'likes', 'dislikes', y 'views' y estás utilizando una regresión lineal múltiple para modelar los datos.

	category_id	views	likes	dislikes	comment_count	publish_day	publish_hour
0	26	7224515.0	55681.0	10247.0	9479.0	5.0	7.0
1	24	1053632.0	25561.0	2294.0	2757.0	7.0	6.0
2	10	17158579.0	787420.0	43420.0	125882.0	5.0	17.0
3	17	27833.0	193.0	12.0	37.0	1.0	2.0
4	25	9815.0	30.0	2.0	30.0	1.0	1.0

3. **Preprocesamiento de Datos**
   Tratar los valores faltantes.
   Aquí un mapa de calor para identificar visualmente los valores faltantes



4. **Construcción del Modelo**
   Obtener conjunto de datos de Entrenamiento (Train) y de Prueba (Test)
   Entrenamos el modelo.

```
numericas= df.select_dtypes(include=['float64', 'int'])
numericas.describe()
```

	views	likes	dislikes	comment_count	publish_day	publish_hour
count	3.862300e+04	3.862300e+04	3.862300e+04	3.862300e+04	38623.000000	38623.000000
mean	5.922344e+06	1.345562e+05	7.059725e+03	1.260477e+04	3.592756	12.945680
std	1.906031e+07	3.503766e+05	4.058532e+04	4.362318e+04	1.739148	6.425201
min	8.510000e+02	0.000000e+00	0.000000e+00	0.000000e+00	1.000000	0.000000
25%	2.456325e+05	5.808500e+03	1.960000e+02	6.620000e+02	2.000000	8.000000
50%	9.758400e+05	2.505200e+04	8.110000e+02	2.452000e+03	4.000000	15.000000
75%	3.672459e+06	1.143460e+05	3.360500e+03	9.182000e+03	5.000000	18.000000
max	4.245389e+08	5.613827e+06	1.753274e+06	1.228655e+06	7.000000	23.000000

5. **Validación del Modelo**
   Evaluar y verificar la precisión, el recall y f1-score utilizando el reporte de clasificación
   Hicimos nuestra matriz de confusión

5. CONCLUSIONES

1. Analizando la imagen 9 de la primera pregunta. Se puede concluir que, las 2 categorías de mayor tendencia son Music y Entertainment. Estas 2 categorías son bastante superiores al resto.

2. Analizando la imagen 10 y 11 de la segunda pregunta. Se puede concluir que, las categorías que más gustan por la cantidad de likes que tienen son Music y Entertainment. Y las categorías que menos gustan son Shows y Travel & Events.

3. Analizando la imagen 12 de la tercera pregunta. Se puede concluir que, las categorías de Shows y Pets & Animals son las que mejor ratio “Me gusta” / “No me gusta” tienen.

4. Analizando la imagen 13 de la cuarta pregunta. Se puede concluir que, las categorías de Travel & Events y Comedy son las que mejor ratio “Vistas” / “Comentarios” tienen.

5. Analizando la imagen 14 de la quinta pregunta. Se puede concluir que, desde el 2017-12 hasta el 2018-03 se ha estado manteniendo el volumen de videos en tendencia. Luego de esto en 2018-04 en adelante se ha visto varios cambios bruscos y el pico más bajo.

6. Analizando la imagen 15 y 16 de la sexta pregunta. Se puede concluir que, los canales de Youtube más frecuentes en la Lista de Videos de tendencia son: “the tonight show with jimmy fallon”, “TheEllenShow” y “Jimmy Kimmel Live”. Y que los canales de Youtube menos frecuentes en la Lista de Videos de tendencia son: “joanjetsetter”, “FPFutebolOficial”, “The Chainsmokers”, entre otros más.

7. Analizando la imagen 17, 18 y 19 de la séptima pregunta. Se puede concluir que, los 5 estados con más vistas en Gran Bretaña son: “Southampton”, “Aberdeen”, “Stockton On Tees”, “South Yorkshire” y “West Midlands”. También que los 5 estados con más me gusta en Gran Bretaña son: “South Yorkshire”, “West Midlands”, “Leicester”, “Devon” y Perthshire And Kinross”. Y que los 5 estados con más no me gusta en Gran Bretaña son: “Bath and North East Somerset ”, “Southampton”, “South Yorkshire”, “Inverclyde” y “Glasgow”.
8. Del modelo de progresión lineal donde utilizamos los las variable ‘likes’, ‘dislikes’, ‘views’ sacamos el siguiente análisis
   Interpretación de los coeficientes:
   Manteniendo todas las demás características fijas, el aumento de 1 Like está asociado con el aumento de 42.251311 vistas.
   Manteniendo todas las demás características fijas, el aumento de 1 Dislike está asociado con el aumento de 38.824725 vistas

6. ARCHIVAR Y COMUNICAR/PUBLICAR:

Enlace al repositorio del trabajo: https://github.com/JhairLVS/FDS2023-2-CC52

7. BIBLIOGRAFÍA