Universidad Peruana de Ciencias Aplicadas

Curso: Fundamentos de data science

Sección: CC52

Nombre del profesor: Nérida Isabel Manrique Tunque

"Informe TF"

Integrantes:

-Jhonny Elias Ruiz Santos

-Franck Manuel Goñas Lopez

-Carlos Daniel Llanos Llamoca

2023-2

ÍNDICE:

I. INTRODUCCIÓN:

Actualmente, el análisis de datos se ha transformado en un pilar fundamental para la toma de decisiones estratégicas en cualquier ámbito empresarial. En el ámbito del marketing digital, se han desarrollado técnicas o métodos con la capacidad de extraer conocimiento significativo de los datos. La finalidad de estos procesos es poder recopilar la mayoría de información de los datos.

El conjunto de datos "Tendencias de las estadísticas de videos de Youtube" se pudo recopilar mediante los diversos factores de medición que utiliza Youtube(número de vistas, compartidos, comentarios y me gusta). Además, la base de datos representa una ventana única hacia la comprensión de las dinámicas y preferencias de la audiencia en una de las plataformas con mayor contenido existente. El registro detallado y diario de los videos más populares en el país de Canada(CA) nos servirá como punto de partida para un proyecto de análisis de datos revelador y detallado.

El objetivo principal de este proyecto es poder realizar una análisis más allá de la superficie de popularidad y la tendencia de los videos. Se trata de comprender las motivaciones de la audiencia, los patrones de consumo para la tendencia de los vídeos, analizar las reacciones y comentarios. Los factores que pueden definir el éxito de un video en un ecosistema digital. La aplicación de modelos de datos será el núcleo de este proyecto, siendo la herramienta más importante para realizar la extracción profunda y significativa de los datos. Esta información identificará similitud y diferencias entre los datos, brindándonos una visión global y estratégica para las decisiones de marketing.

En resumen, este proyecto de análisis de datos sobre las tendencias de los videos de Youtube busca transformar los datos crudos en información estratégica. Se trata de proporcionar a una empresa de marketing digital las respuestas necesarias para optimizar sus estrategias competitivas, aumentar su alcance global y maximizar el impacto en la plataforma, por el que se necesitará obtener todos los datos necesarios ya que la compresión de la audiencia es esencial para el éxito de los videos.

II. Roles de los integrantes

Rol	Integrante	Tareas Asignadas
Business Project Sponsor	Carlos Daniel Llanos Llamoca	Realizar la parte informativa del proyecto, como la redacción del informe y describir los resultados
Data Science	Jhonny Elias Ruiz Santos	Realizar todo el proceso de preparación y limpieza de dato
Data Engineer / Data Analytics	Franck Manuel Goñas Lopez	Realizar los gráficos, tablas y modelos necesarios para dar respuesta a las problemáticas.

III. METODOLOGÍA CRISP-DM

1. *Comprensión del negocio:*

Objetivos del proyecto:

- Analizar los patrones y comportamientos de los videos más populares de Canadá, extrayendo datos sobre duración de visualización, interacciones, reacciones del público y métricas relevantes que definen la popularidad de los videos.
- Identificar las preferencias y diferencias entre los diversos datos del país de Canadá en cuanto a géneros de videos, temas específicos y características que generan mayor visualización.

- Determinar las variables que tendrán una correlación significativa con la popularidad y los éxitos de los videos en Youtube.

Objetivos de Data Science:

- **Variable dependiente:** Popularidad del video(métricas obtenidas como el número de likes, vistas, dislikes, comentarios, etc.)
- **Variable Independiente:** Duración del video, categoría, interacciones sociales, hora de publicaciones, entre otros tipos de conjuntos relevantes.

2. *Comprensión de los datos:*

Los tipos de datos son los siguientes:

```
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   video_id               40881 non-null  object
 1   trending_date          40881 non-null  object
 2   title                  40881 non-null  object
 3   channel_title          40881 non-null  object
 4   category_id            40881 non-null  int64
 5   publish_time           40881 non-null  object
 6   tags                   40881 non-null  object
 7   views                  40881 non-null  int64
 8   likes                  40881 non-null  int64
 9   dislikes               40881 non-null  int64
 10  comment_count          40881 non-null  int64
 11  thumbnail_link         40881 non-null  object
 12  comments_disabled      40881 non-null  bool
 13  ratings_disabled       40881 non-null  bool
 14  video_error_or_removed 40881 non-null  bool
 15  description            39585 non-null  object
 16  state                  40881 non-null  object
 17  lat                    40881 non-null  float64
 18  lon                    40881 non-null  float64
 19  geometry               40881 non-null  object
dtypes: bool(3), float64(2), int64(5), object(10)
```

Figura 1: Visualización de los datos

Se puede visualizar que existen un total de 20 tipos de datos en nuestra dataset, nos podemos asegurar de visualizar correctamente el total de variables para realizar un análisis exitoso de los datos.

3. *Preparación de los datos:*

```
dataset = dataset.dropna()
```

```
valores_nulos = dataset.isnull().sum()
print(valores_nulos)
```

```
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               0
state                     0
lat                       0
lon                       0
geometry                  0
dtype: int64
```

Para la preparación de los datos usamos la función dropna la cual nos permite eliminar todos los valores nulos de nuestro dataset.

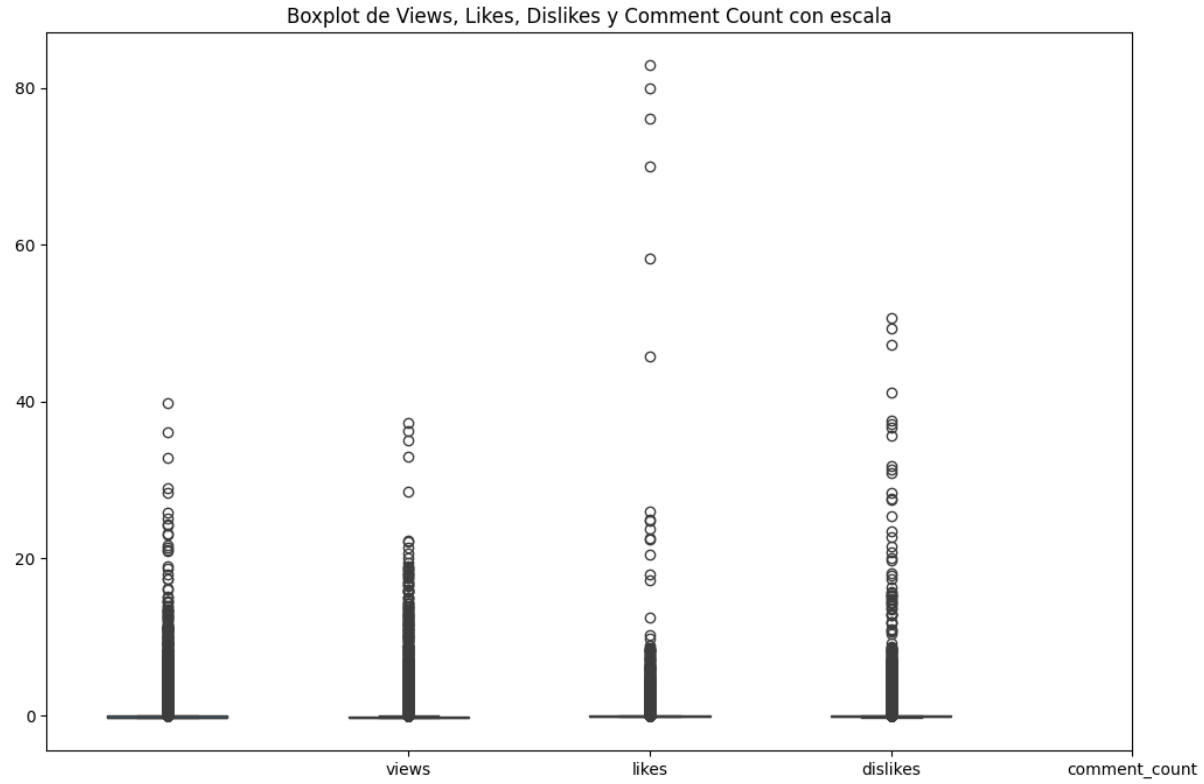
```
import json

id_to_category = {}

with open(r'../data/CA_category_id.json') as file:
    js = json.load(file)

for category in js["items"]:
    id_to_category[category["id"]] = category["snippet"]["title"]
dataset["category"] = dataset["category_id"].map(id_to_category)
```

Asimismo, se creó una columna uniendo el json al dataset para así tener el nombre de la categoría.



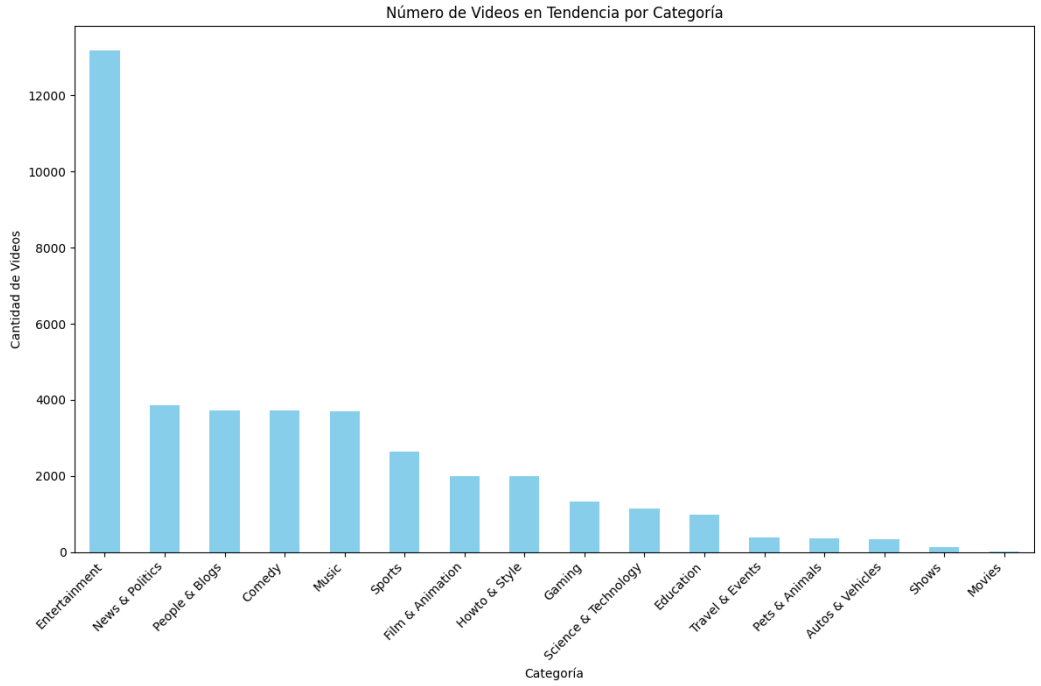
Se revisaron los elementos outliers en views, likes, dislikes y comment count. La idea principal era reemplazarlos por la media pero al darnos cuenta de que la cantidad de outliers era representativa, no lo hicimos.

```
Variable 'views': 0.10852595680181887 outliers
Variable 'likes': 0.12295061260578502 outliers
Variable 'dislikes': 0.13045345459138563 outliers
Variable 'comment_count': 0.11933813313123658 outliers
```

Por otro lado, para el caso de la categoría usamos la función json.load para poner cargar el archivo json que incluye las categorías, luego lo relacionamos con el category_id del dataset para poder unirlos y saber la categoría de cada video

Pregunta 1:

¿Qué categorías de videos son las de mayor tendencia?

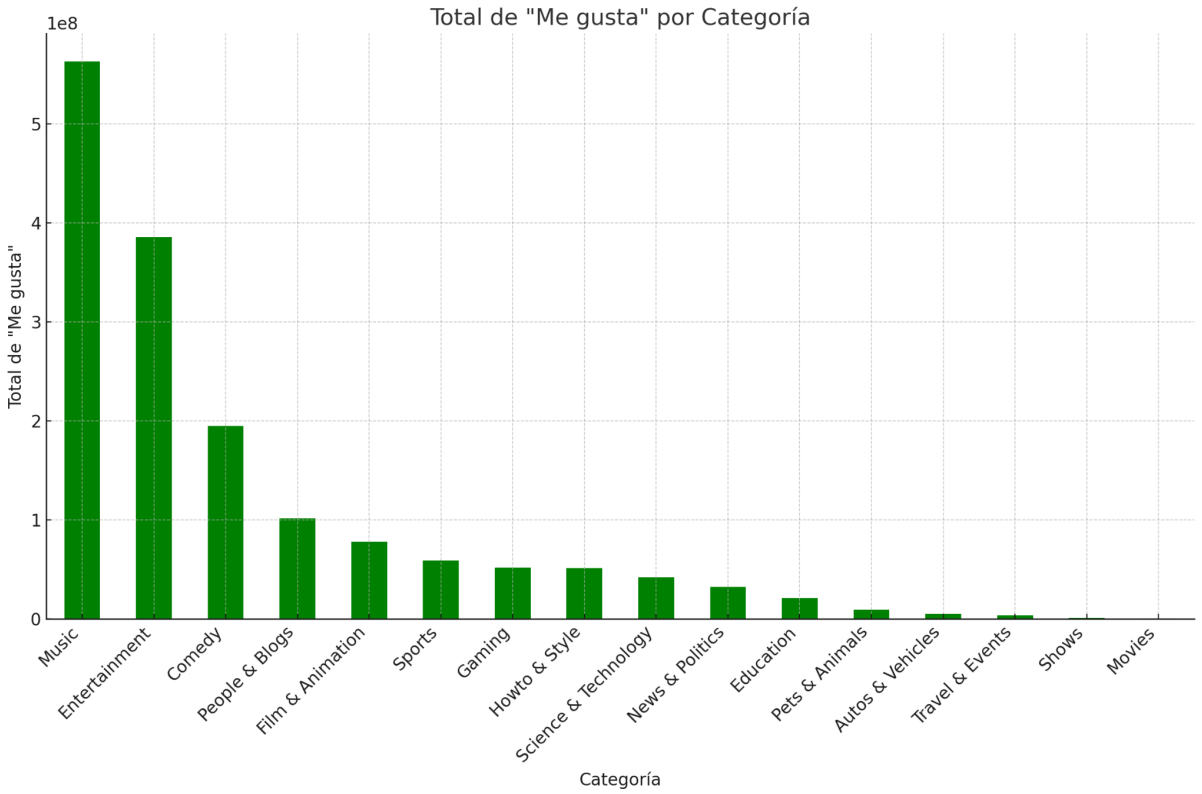


	Category	Number of Videos
0	Entertainment	13173
1	News & Politics	3868
2	People & Blogs	3726
3	Comedy	3725
4	Music	3695
5	Sports	2650
6	Film & Animation	2001
7	Howto & Style	1998
8	Gaming	1330
9	Science & Technology	1143

Como podemos visualizar en el gráfico y en el listado de las categorías con mayor tendencia en el país de Canadá. La categoría de videos con mayor tendencia es de “Entretenimiento” mientras que los videos de menor tendencia son los de categoría “Películas”. Se espera un aumento de la cantidad de los videos de categoría de “Entretenimiento” para un mayor aumento en la atención de los usuarios.

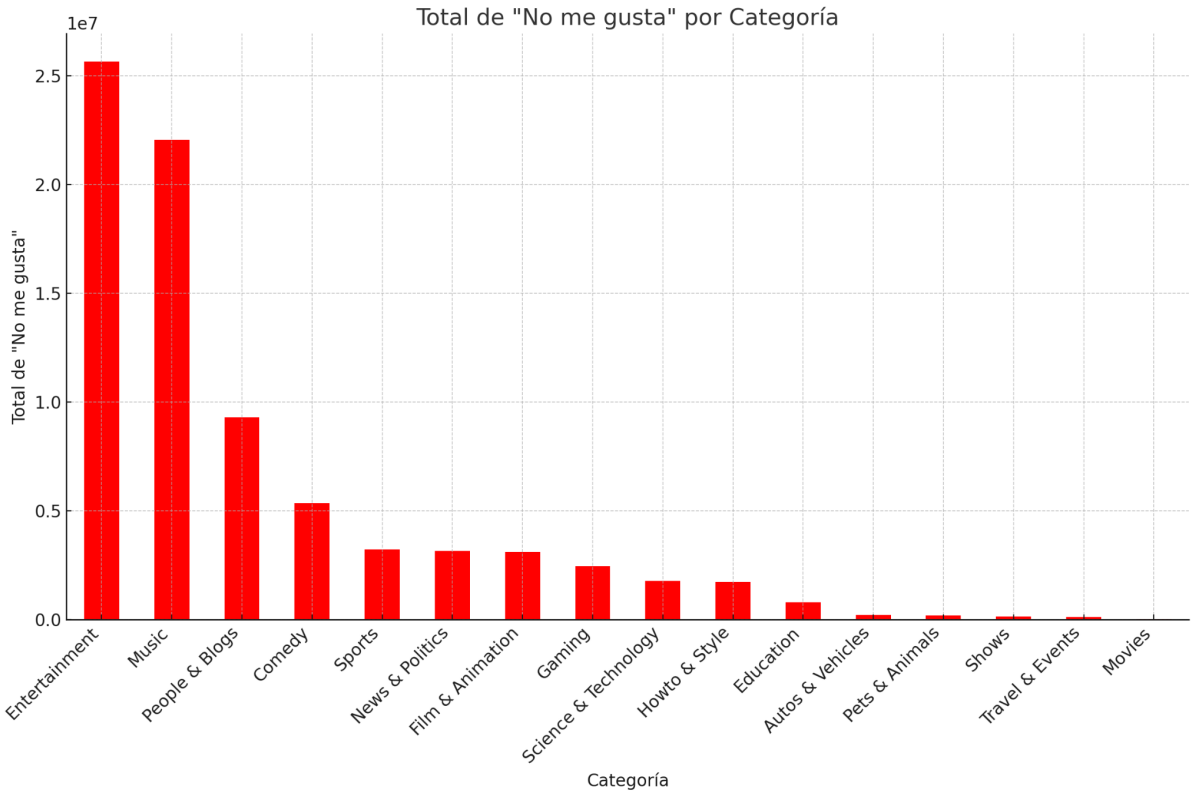
Pregunta 2:

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



	Category	Total Likes
0	Music	562950999
1	Entertainment	385612477
2	Comedy	194794035
3	People & Blogs	101413106
4	Film & Animation	77693165
5	Sports	58574334
6	Gaming	51538080
7	Howto & Style	51149090
8	Science & Technology	42013552
9	News & Politics	32202728

Si visualizamos el gráfico, podemos ver que las categorías de Música con un total 562M de likes y Entretenimiento con un total de 385M de likes poseen la mayor cantidad de “me gusta” entre todas las categorías según la recopilación de datos en la región de Canadá.



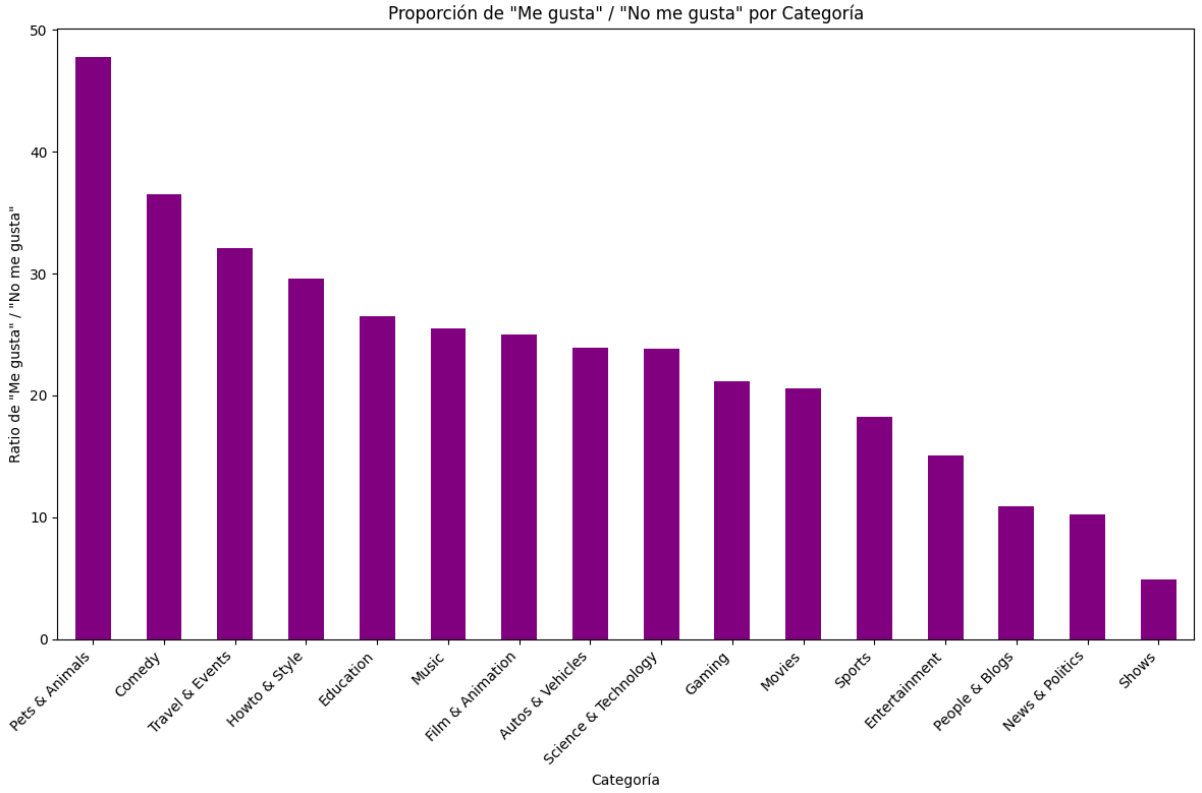
	Category	Total Dislikes
0	Entertainment	25660825
1	Music	22064769
2	People & Blogs	9296718
3	Comedy	5341733
4	Sports	3209078
5	News & Politics	3146614
6	Film & Animation	3104400
7	Gaming	2434569
8	Science & Technology	1765944
9	Howto & Style	1727377

Si visualizamos el gráfico, podemos ver que las categorías de Entretenimiento con un total 25M de dislikes y Música con un total de 22M de dislikes poseen la mayor cantidad de “no me gusta” entre todas las categorías según la recopilación de datos en la región de Canadá.

En conclusión, las categorías de Entretenimiento y Música poseen la mayor cantidad de likes y dislikes entre todas las categorías en el país de Canadá.

Pregunta 3:

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

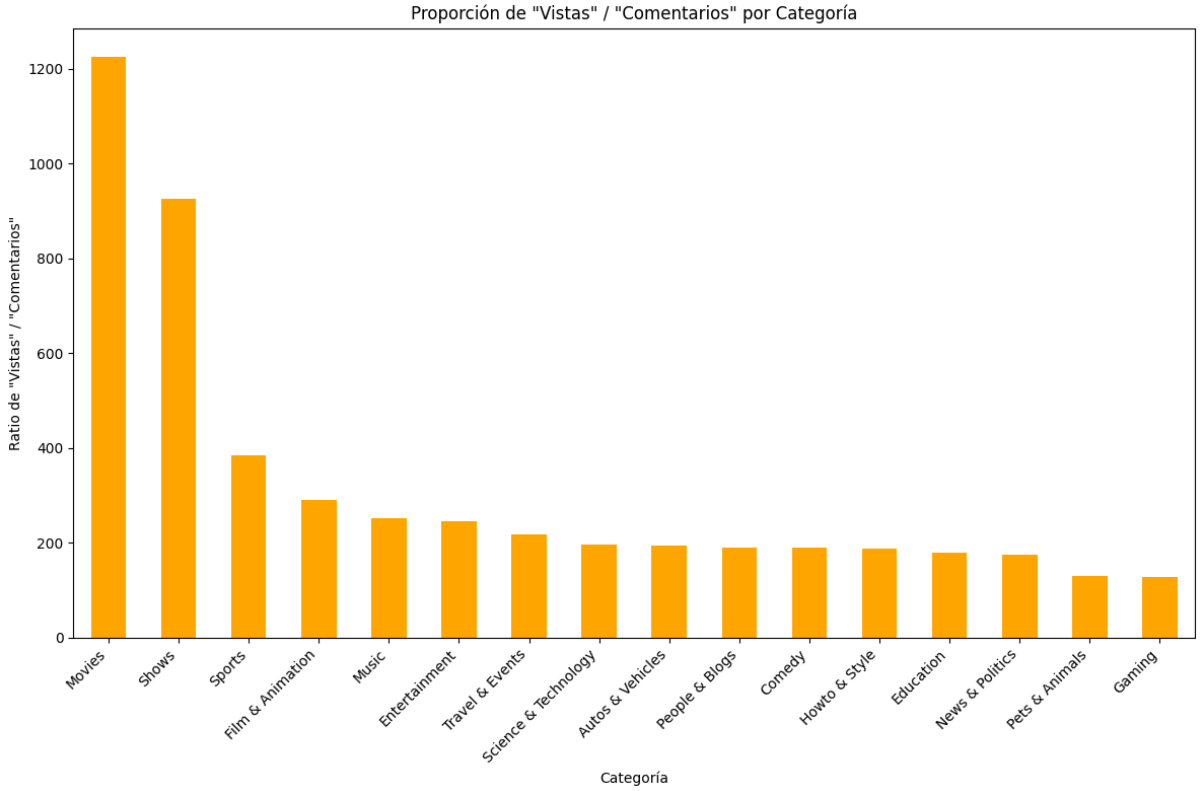


	Category	Like/Dislike Ratio
0	Pets & Animals	47.732676
1	Comedy	36.466442
2	Travel & Events	32.081307
3	Howto & Style	29.610826
4	Education	26.475243
5	Music	25.513568
6	Film & Animation	25.026781
7	Autos & Vehicles	23.928657
8	Science & Technology	23.790974
9	Gaming	21.169274

Las categorías con las mejores proporciones(like/dislike) son las de “Mascotas & Animales” , “Comedia” y “Viajes & Eventos”. Estas categorías se encuentran ubicadas en la cúspide del listado con respecto al radio de likes/dislikes entre el rango de (30-50)% en el país de Canadá.

Pregunta 4:

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

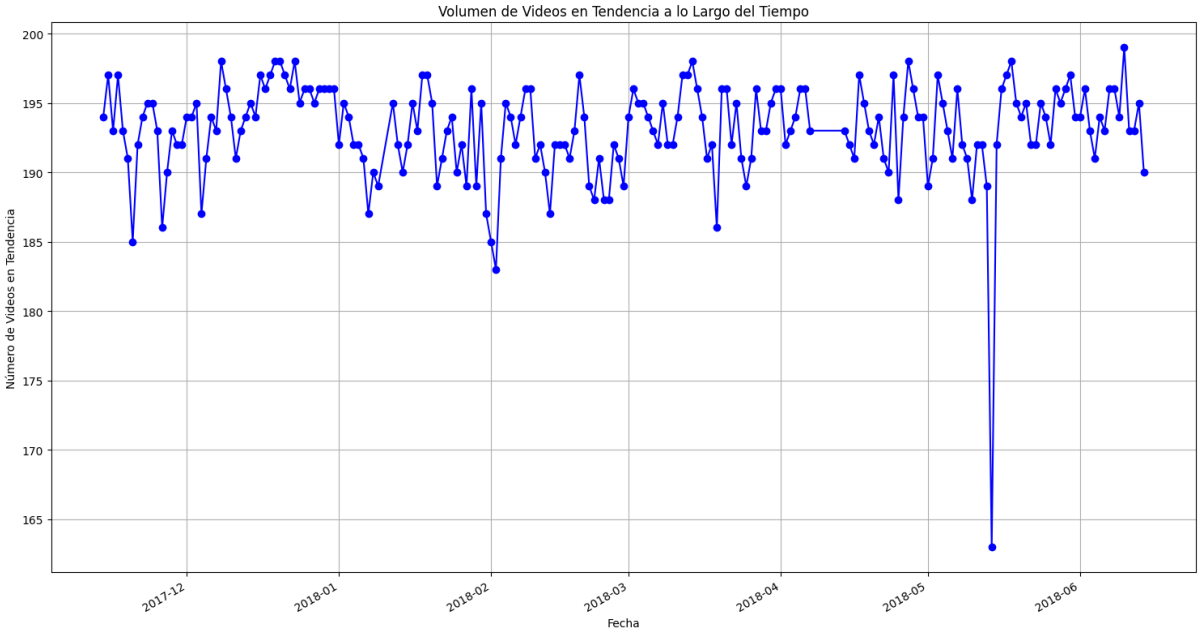


	Category	View/Comment Ratio
0	Movies	1224.641631
1	Shows	926.279167
2	Sports	384.231775
3	Film & Animation	289.976372
4	Music	251.325162
5	Entertainment	245.975988
6	Travel & Events	217.767645
7	Science & Technology	196.660274
8	Autos & Vehicles	194.808564
9	People & Blogs	190.026795

Como se visualiza en el gráfico, la categoría con el mayor ratio de vistas/comentarios es “Películas” . La siguiente categoría con un mayor ratio de vistas/comentarios es “Shows”. Estas dos categorías son las de mayor ratio a comparación de las otras categorías en Canadá.

Pregunta 5:

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

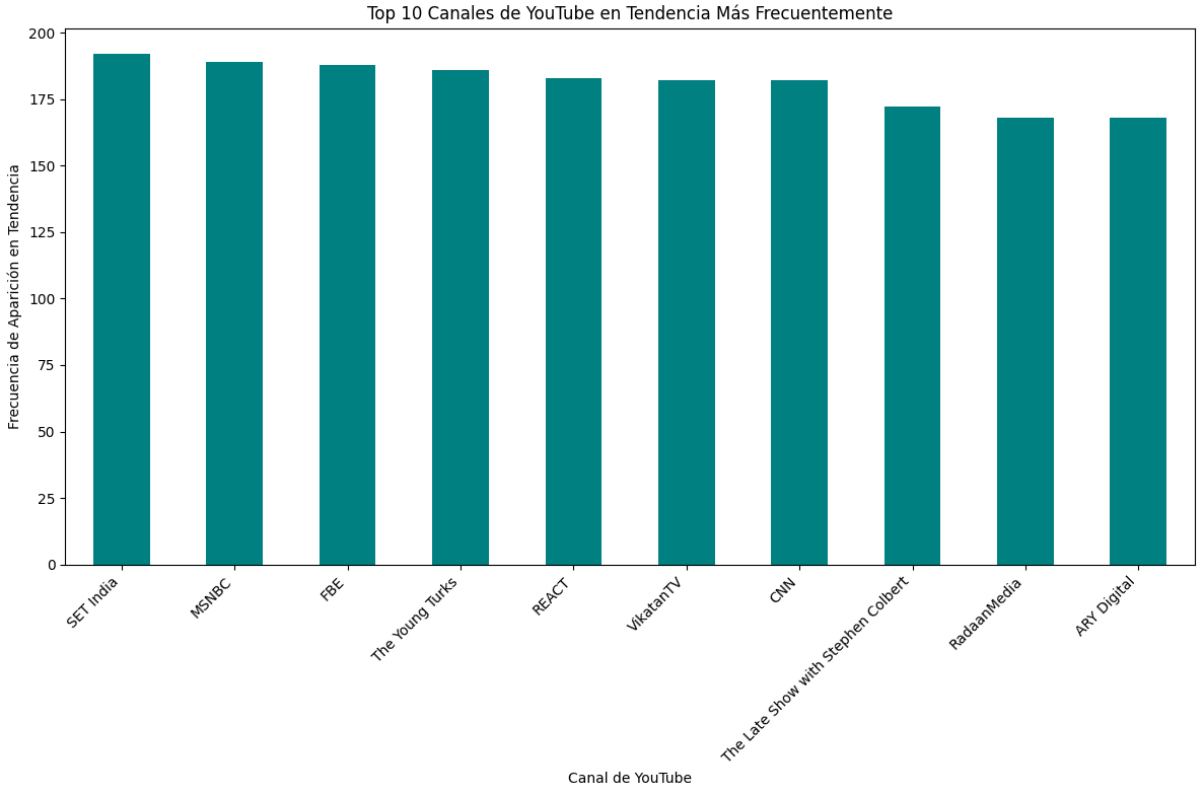


	Trending Date	Number of Trending Videos
0	2017-11-14	194
1	2017-11-15	197
2	2017-11-16	193
3	2017-11-17	197
4	2017-11-18	193
5	2017-11-19	191
6	2017-11-20	185
7	2017-11-21	192
8	2017-11-22	194
9	2017-11-23	195

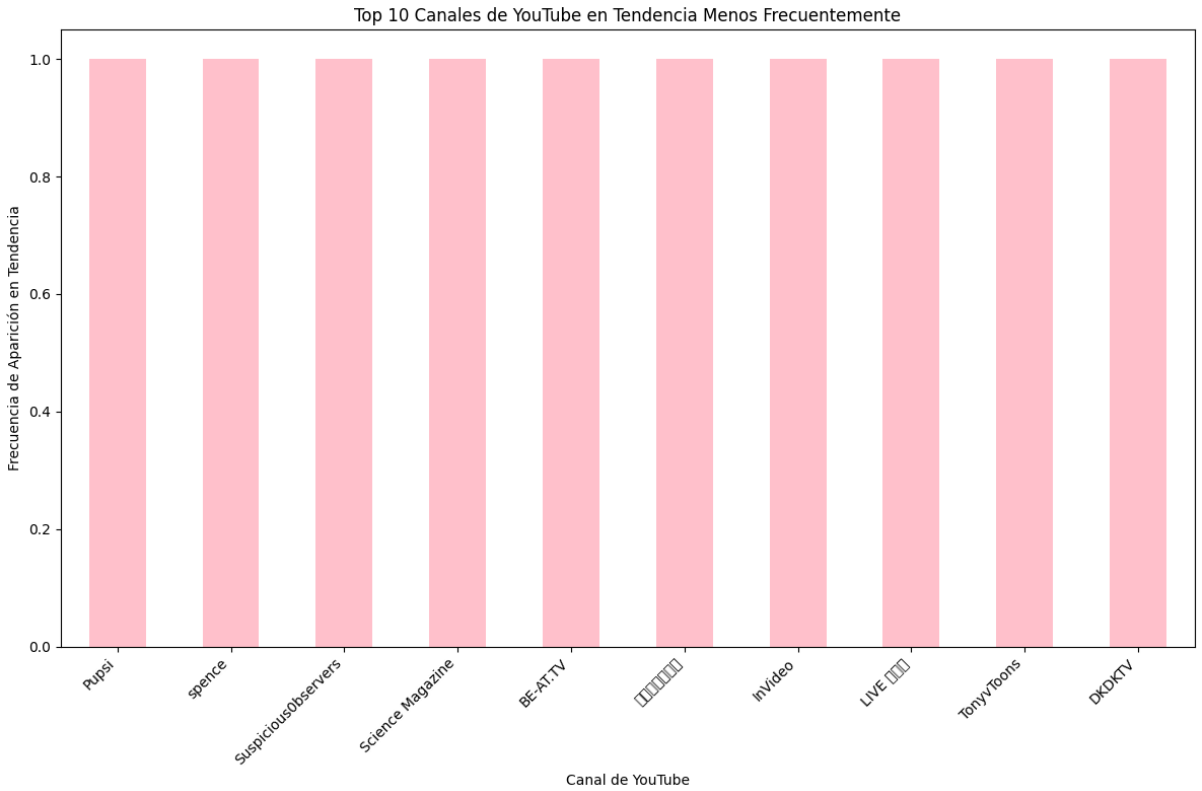
Mediante el gráfico, se puede evidenciar un volumen de videos en tendencia constante ubicado entre un rango de 180-200 en la región de Canadá(CA). Este rango se pudo obtener mediante la información recopilada entre los años 2017 y 2018.

Pregunta 6:

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



	Channel Title	Trending Frequency
0	SET India	192
1	MSNBC	189
2	FBE	188
3	The Young Turks	186
4	REACT	183
5	VikatanTV	182
6	CNN	182
7	The Late Show with Stephen Colbert	172
8	RadaanMedia	168
9	ARY Digital	168

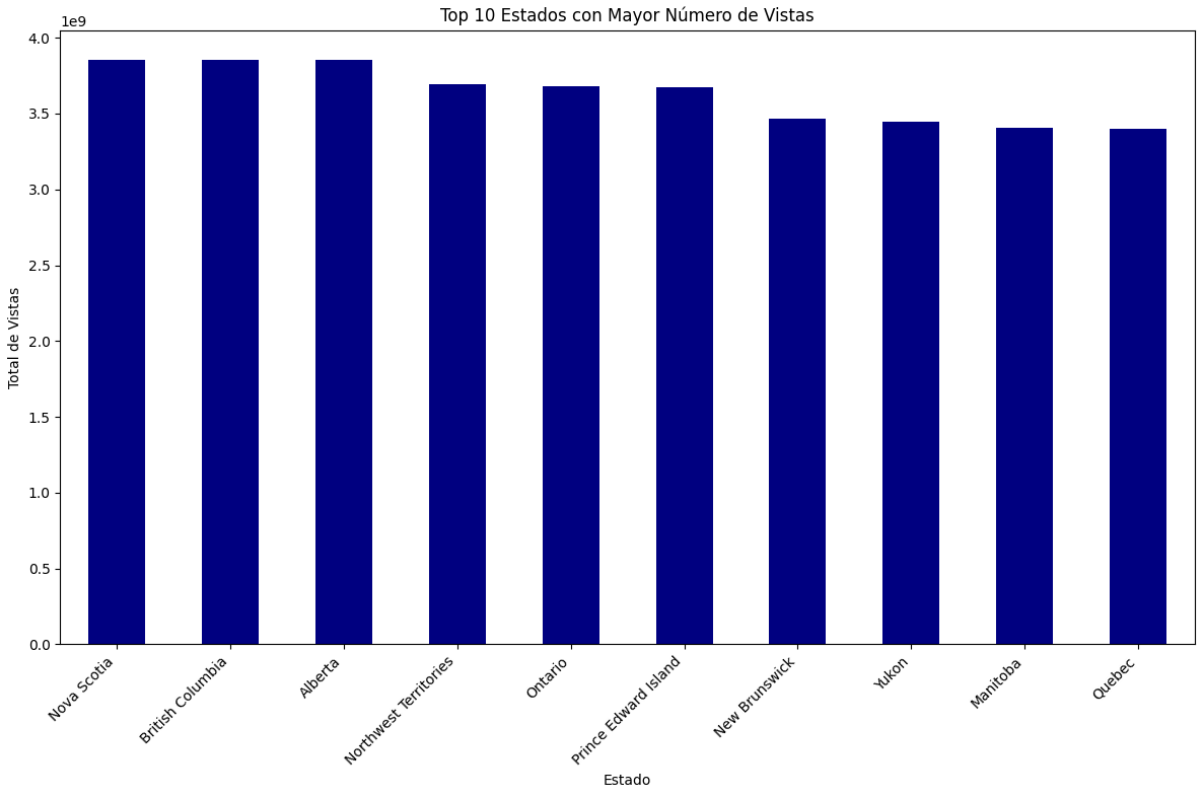


	Channel Title	Trending Frequency
0	Pupsi	1
1	spence	1
2	Suspicious0bservers	1
3	Science Magazine	1
4	BE-AT.TV	1
5	中国有嘻嘻哈哈	1
6	InVideo	1
7	LIVE 郭文贵	1
8	TonyvToons	1
9	DKDKTV	1

Mediante los gráficos y listados realizados por la recopilación de datos sobre los canales. Se pudo evidenciar que los canales con mayor frecuencia en Canadá son SET India, MSNBC y FBE. Mientras que los canales con menor frecuencia son Pupsi, Spence, entre otros.

Pregunta 7:

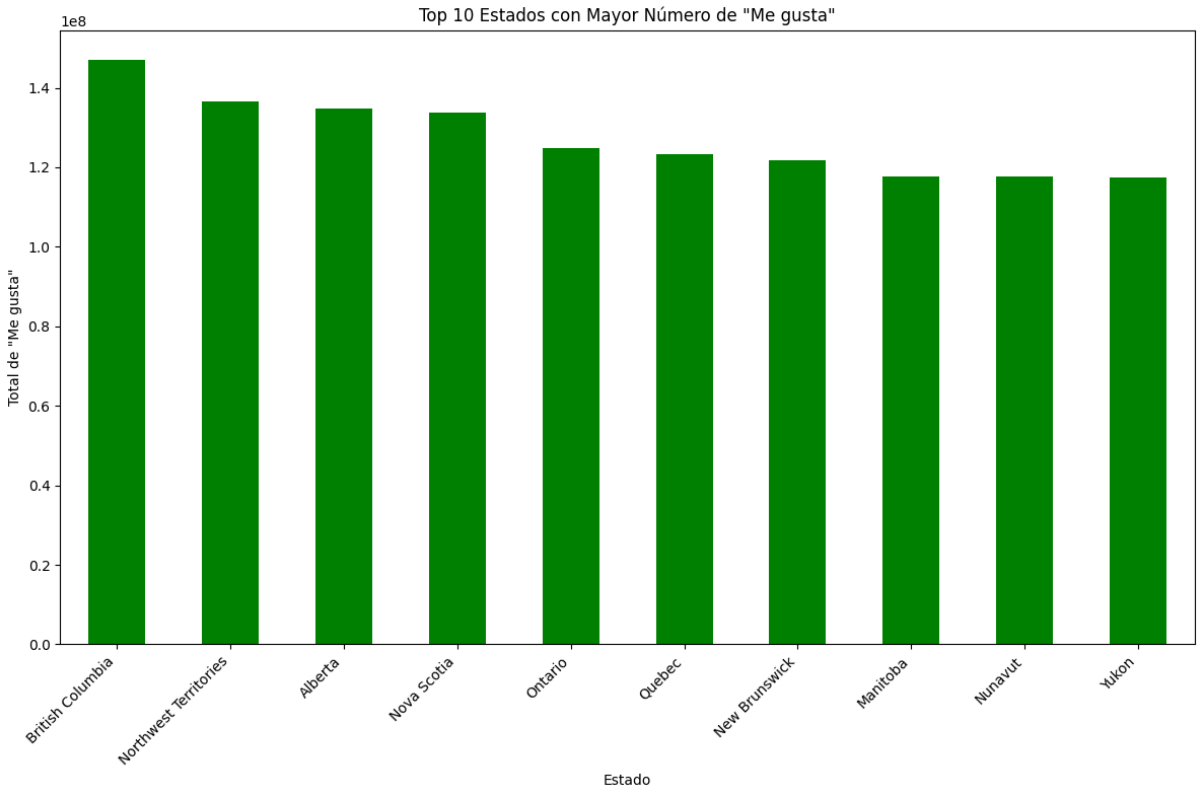
¿En qué Estados se presenta el mayor número de “Vistas”, “Me gusta” y “No me gusta”?



findfont: Generic family 'sans-serif' n

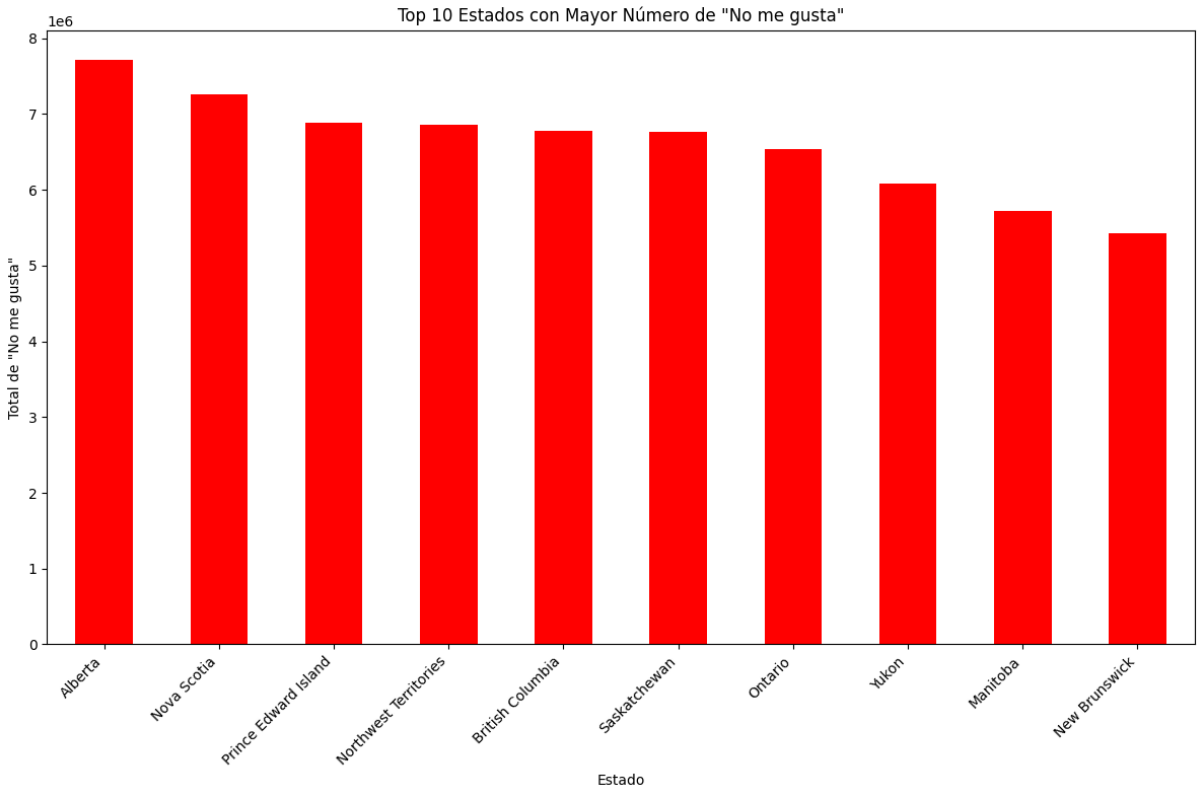
	State	Total Views
0	Nova Scotia	3856335841
1	British Columbia	3853586604
2	Alberta	3851814774
3	Northwest Territories	3690688489
4	Ontario	3682550609
5	Prince Edward Island	3672850678
6	New Brunswick	3469101073
7	Yukon	3447525858
8	Manitoba	3404276252
9	Quebec	3396978811

Como podemos visualizar en el gráfico, el estado con mayor número de vistas es Nova Scotia con un alrededor de 3.8B de vistas en el rango de tiempo determinado.



	State	Total Likes
0	British Columbia	147110449
1	Northwest Territories	136638219
2	Alberta	134910354
3	Nova Scotia	133693555
4	Ontario	124856852
5	Quebec	123292815
6	New Brunswick	121655482
7	Manitoba	117818126
8	Nunavut	117681262
9	Yukon	117444017

Mediante el listado de los estados con mayor número de likes en los videos de Youtube, se puede visualizar que el estado con mayor número de likes otorgado por sus usuarios es British Columbia.



	State	Total Dislikes
0	Alberta	7719223
1	Nova Scotia	7262751
2	Prince Edward Island	6891081
3	Northwest Territories	6863719
4	British Columbia	6783569
5	Saskatchewan	6770597
6	Ontario	6542266
7	Yukon	6081053
8	Manitoba	5721698
9	New Brunswick	5422132

Se puede visualizar mediante el gráfico, los habitantes del estado de Alberta son los usuarios que otorgan el mayor número de dislikes en los videos de Youtube..

4. Modelizar y Evaluar resultados:

Pregunta 8:

¿Es factible predecir el número de “Vistas” o “Me gusta” o “No me gusta”? Si es posible para ello se va crear un modelo de regresión lineal múltiple para predecir los me gusta, teniendo en cuenta la cantidad de vistas, los no me gusta y la cantidad de comentarios. El modelo tiene una precisión de 0.8547

```
dataset = pd.read_csv('dataset.csv')


features = dataset[['views', 'dislikes', 'comment_count']]  # características
target = dataset['likes']  # Variable objetivo

# Dividir el conjunto de datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.3, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

predictions = model.predict(X_test)

# Evaluar el modelo
mse = mean_squared_error(y_test, predictions)
r2 = r2_score(y_test, predictions)

print(f"Error cuadrático medio (MSE): {mse}")
print(f"Coeficiente de determinación (R^2): {r2}")
```

```
[48]  ✓ 1.2s

...  Error cuadrático medio (MSE): 3142856508.2512145
     Coeficiente de determinación (R^2): 0.854733485277898
```

Pregunta 9:

¿Los videos en tendencia son los que mayor cantidad de comentarios positivos reciben?

Para poder responder a esa pregunta primero tendríamos que saber cuando un comentario es positivo, para ello se podría utilizar el análisis de sentimientos o un análisis de regresión logística que nos permitiría modelar 0 o 1 según si el comentario es positivo o negativo. Sin embargo, en el dataset no se dispone de los comentarios.

IV. CONCLUSIONES

- Se logró un entendimiento profundo de los factores que influyen en la popularidad de los videos en YouTube, como la duración, interacciones y reacciones del público.
- Las categorías de "Entretenimiento" y "Música" dominan en popularidad, generando las mayores reacciones positivas y negativas.
- Aspectos como la duración del video y la interacción social son claves en la popularidad de los contenidos en YouTube.
- La limpieza y organización de los datos, incluyendo la decisión de no alterar los outliers, fueron esenciales para un análisis preciso.
- Categorías como "Mascotas & Animales" y "Comedia" destacan por su buena recepción y altas proporciones de reacciones positivas.
- Se observaron diferencias regionales en Canadá en cuanto a visualizaciones y "me gusta", indicando variaciones en las preferencias del público.
- Se desarrolló un modelo de regresión lineal múltiple eficaz para predecir la popularidad de los videos basándose en varias métricas.
- Se sugiere la futura utilización de análisis de sentimientos para explorar la relación entre comentarios y popularidad de los videos.