Week 2 Ingest and Explore the Dataset

Team: Wendi Yuan, Yujia Cao, Yuhan Zhao

1．What are the predictors and why?

We live in the era of big data, as is widely known. YouTube is no longer just a streaming platform; it has grown into an ecosystem that is deeply ingrained into our daily lives. From social interactions to learning, entertainment, and lifestyle, YouTube plays a major role. Through our analysis of data, we aim to understand what drives YouTube video popularity from the perspectives of YouTube, advertisers, and content creators.

Our goal is to develop a system that can track real-time data to identify the latest trending channels and content. Using this method, we will be able to keep track of creators' activities and provide an up-to-date insight into new content. We will sort the numerical data, first analyzing the latest and most popular trends by ranking them. Then, using this ranking, we can identify and filter out the most trending keywords. Natural language processing can be used to generate multidimensional ranking lists, perform sentiment analysis, and discover trending topics in short videos.

In addition, we plan to train machine learning algorithms, such as RNNs, to generate YouTube comments. By doing this, advertisers will be able to identify next month's trending keywords to improve ad targeting, as well as assist the platform with better traffic distribution, and creators will be able to boost their views.

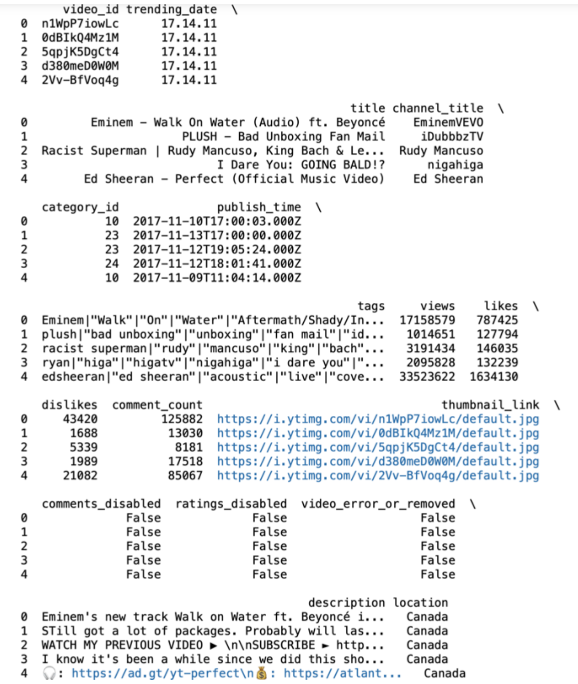
2. What is the target variable and why?

Our goal this week is to analyze data from five countries: Canada, Germany, France, Great Britain, and the USA. Our dataset consists of 16 variables, and we have selected four key numerical metrics—views, likes, dislikes, and comment count—for our analysis. We aim to identify the factors that most significantly impact viewership. That will help us analyzing the latest and most popular trends by ranking them.

Additionally, we have chosen to focus on the title, tags, and description for natural language processing. Through sentiment analysis, we hope to uncover trending patterns and key terms to aid in training future language models.

3. Exploration of the dataset: definition of variables, data types, general dataset stats: count of rows, count of columns, etc.

The dataset consists of 17 variables, including key video attributes such as video_id, title, channel_title, and category_id, alongside engagement metrics like views, likes, dislikes, and comment_count. Additionally, it provides metadata, including the publish_time, tags, thumbnail_link, and whether comments or ratings are disabled (comments_disabled, ratings_disabled). Boolean fields also indicate whether a video has encountered an error or been removed (video_error_or_removed). The dataset contains a mix of data types, with text fields for video descriptions and tags, integers for metrics like views and likes, booleans for the disabled fields, and datetime fields for publish_time and trending_date. This structure provides ample opportunities for both numerical and natural language analysis. The dataset contains 5 rows and thousands of entries, providing rich information for trends, patterns, and outlier detection across multiple variables.

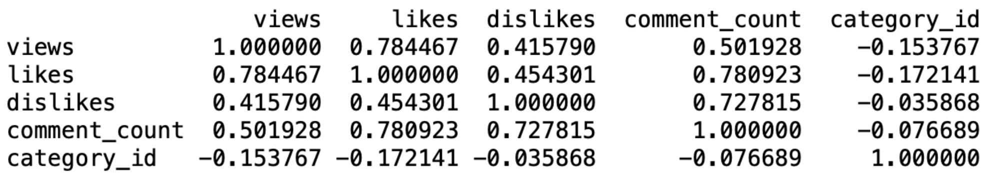


```
      video_id trending_date  \
0  n1WpP7iowLc      17.14.11
1  0dBIkQ4Mz1M      17.14.11
2  5qpjK5DgCt4      17.14.11
3  d380meD0W0M      17.14.11
4  2Vv-BfVoq4g      17.14.11

                                               title channel_title  \
0         Eminem - Walk On Water (Audio) ft. Beyoncé    EminemVEVO
1                     PLUSH - Bad Unboxing Fan Mail     iDubbbzTV
2  Racist Superman | Rudy Mancuso, King Bach & Le...  Rudy Mancuso
3                          I Dare You: GOING BALD!?      nigahiga
4        Ed Sheeran - Perfect (Official Music Video)    Ed Sheeran

   category_id              publish_time  \
0           10  2017-11-10T17:00:03.000Z
1           23  2017-11-13T17:00:00.000Z
2           23  2017-11-12T19:05:24.000Z
3           24  2017-11-12T18:01:41.000Z
4           10  2017-11-09T11:04:14.000Z

                                                tags     views    likes  \
0  Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...  17158579   787425
1  plush|"bad unboxing"|"unboxing"|"fan mail"|"id...   1014651   127794
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...   3191434   146035
3  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...   2095828   132239
4  edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...  33523622  1634130

   dislikes  comment_count                                  thumbnail_link  \
0     43420         125882  https://i.ytimg.com/vi/n1WpP7iowLc/default.jpg
1      1688          13030  https://i.ytimg.com/vi/0dBIkQ4Mz1M/default.jpg
2      5339           8181  https://i.ytimg.com/vi/5qpjK5DgCt4/default.jpg
3      1989          17518  https://i.ytimg.com/vi/d380meD0W0M/default.jpg
4     21082          85067  https://i.ytimg.com/vi/2Vv-BfVoq4g/default.jpg

   comments_disabled  ratings_disabled  video_error_or_removed  \
0              False             False                   False
1              False             False                   False
2              False             False                   False
3              False             False                   False
4              False             False                   False

                                         description location
0  Eminem's new track Walk on Water ft. Beyoncé i...   Canada
1  STill got a lot of packages. Probably will las...   Canada
2  WATCH MY PREVIOUS VIDEO ▶ \n\nSUBSCRIBE ► http...   Canada
3  I know it's been a while since we did this sho...   Canada
4  🎧: https://ad.gt/yt-perfect\n💰: https://atlant...   Canada
```

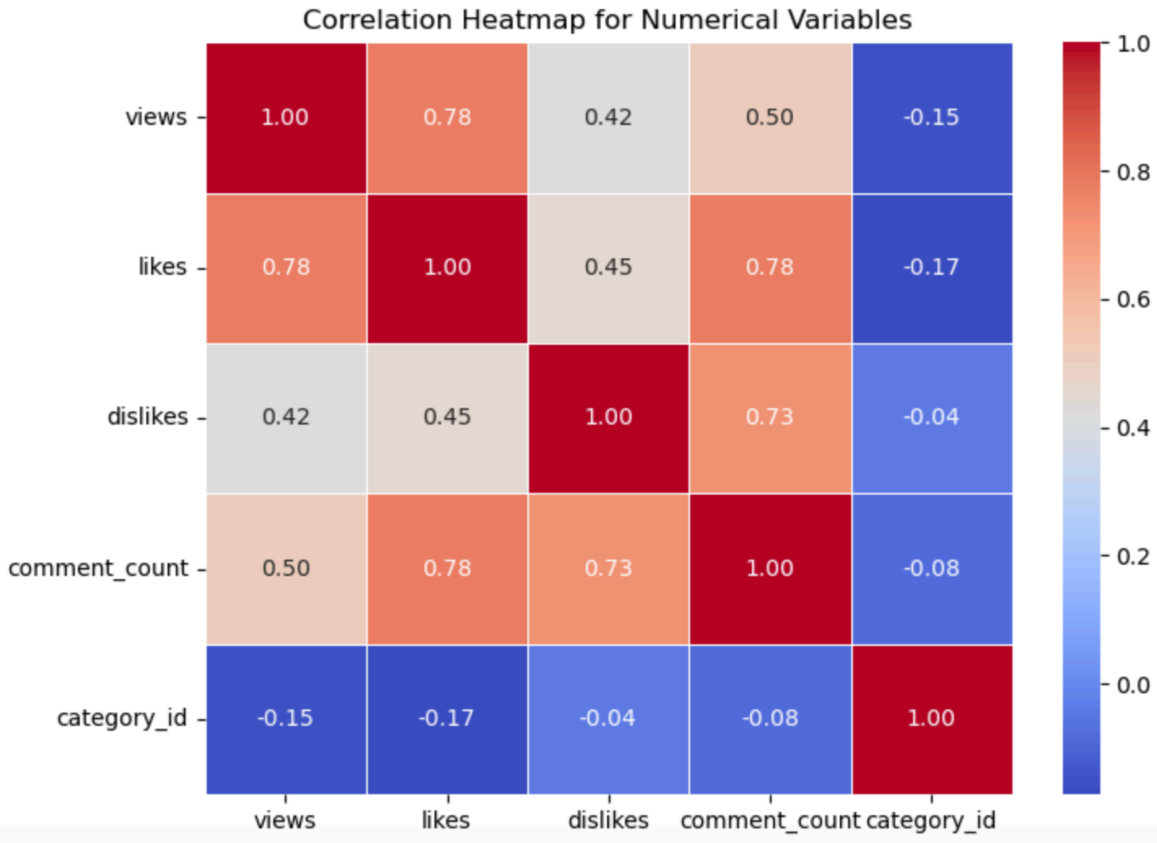
data exploration for numerical variables

The histograms of views, likes, dislikes, and comment count show a clear right-skewness, meaning that most videos have very low engagement. The majority of videos have few views and likes, while only a small number of videos have accumulated hundreds of millions of views or millions of likes. This pattern is common in social media platforms, where only a handful of viral content generates a significant portion of the total engagement, leaving most videos with minimal interaction. Given this skewed distribution, using the mean alone to describe the data may not be effective, as it would be heavily influenced by outliers.

EDA

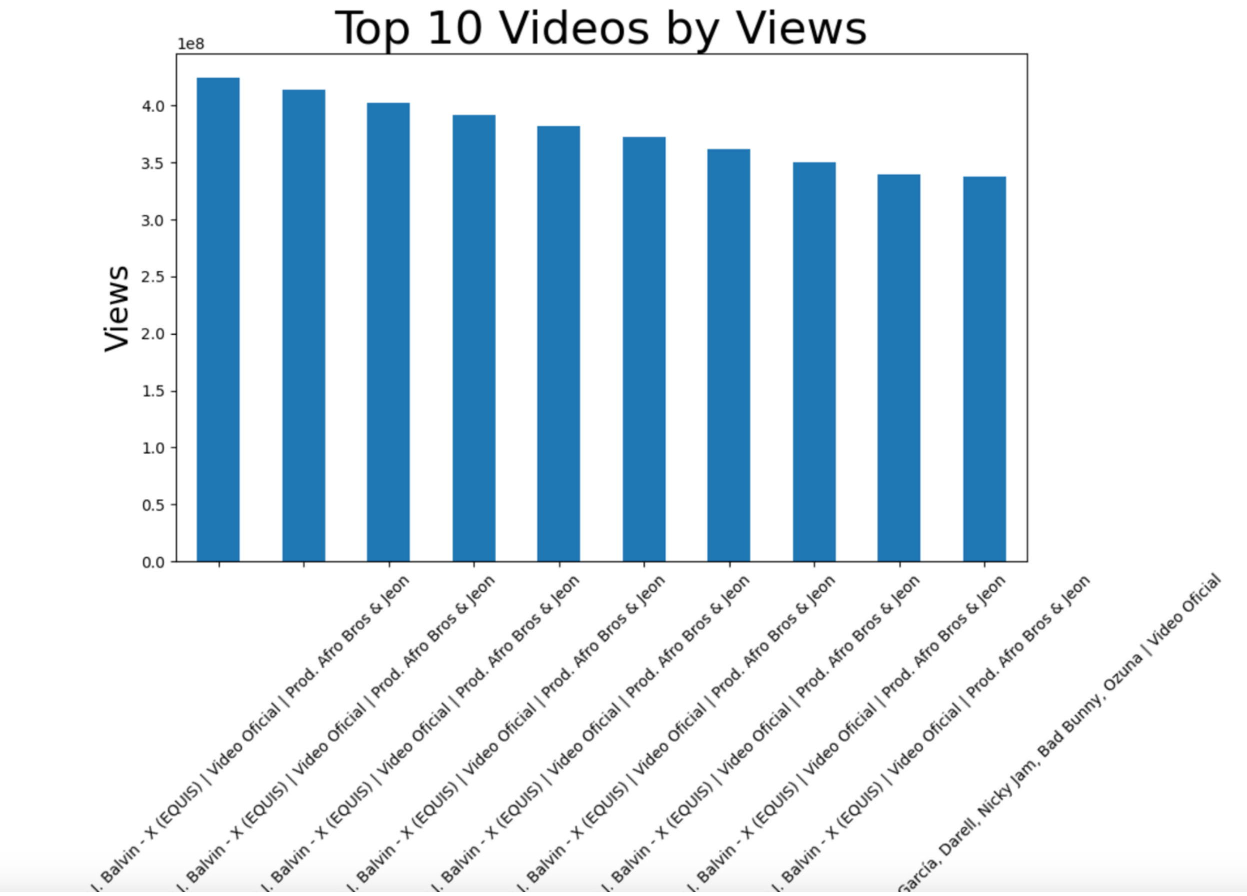


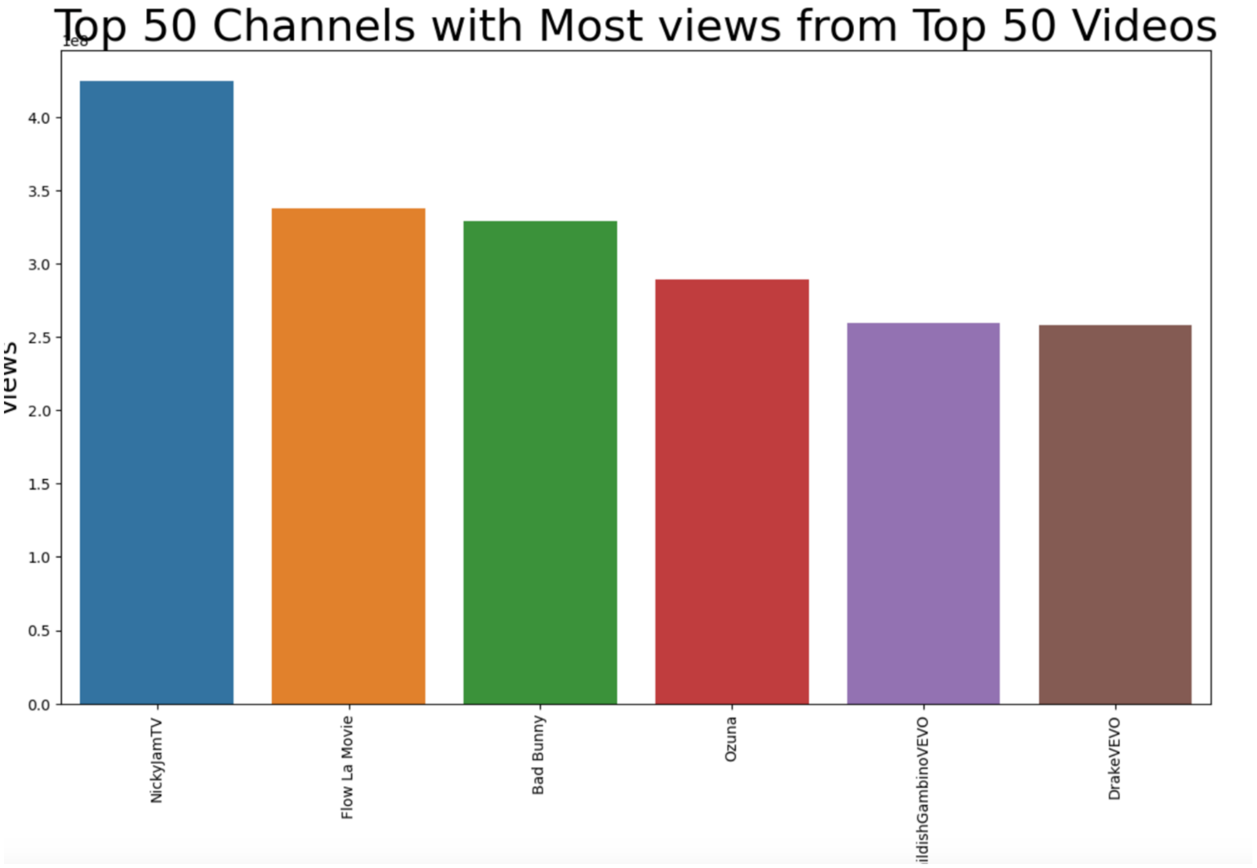
```
                 views     likes  dislikes  comment_count  category_id
views         1.000000  0.784467  0.415790       0.501928    -0.153767
likes         0.784467  1.000000  0.454301       0.780923    -0.172141
dislikes      0.415790  0.454301  1.000000       0.727815    -0.035868
comment_count 0.501928  0.780923  0.727815       1.000000    -0.076689
category_id  -0.153767 -0.172141 -0.035868      -0.076689     1.000000
```

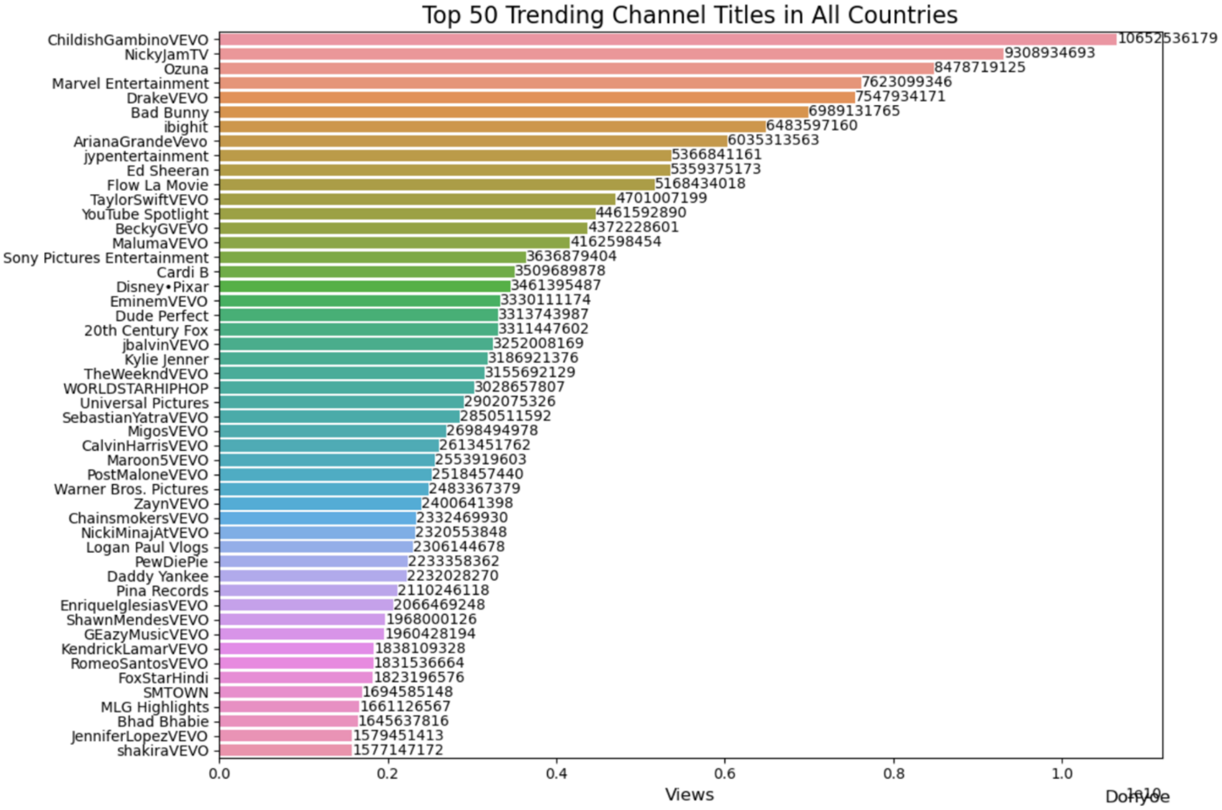


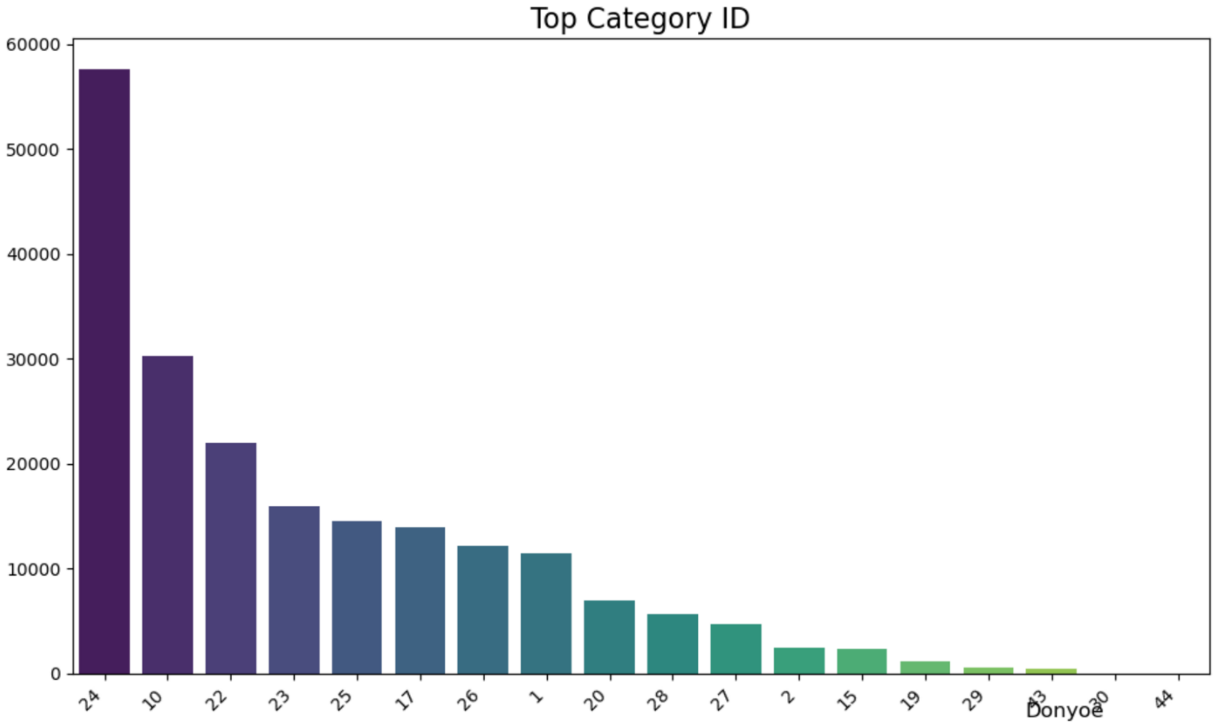
The correlation heatmap reveals strong positive relationships between key engagement metrics. Views have a strong correlation with likes (0.78), indicating that videos with higher view counts tend to receive more likes. Likes are also strongly correlated with comment count (0.78) , comment count and dislikes (0.73), implying that videos generating more discussion tend to be more polarizing. However, category ID shows weak or negative correlations with other variables, indicating that the category a video belongs to has little impact on its engagement metrics.

We are ranking some top trending views base on different variables, which will help us with the upcoming NLP analysis.


Top 10 Videos by Views
1e8
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Views
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
García, Darell, Nicky Jam, Bad Bunny, Ozuna | Video Oficial


Top 50 Channels with Most views from Top 50 Videos
1e8
views
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
NickyJamTV
Flow La Movie
Bad Bunny
Ozuna
ildishGambinoVEVO
DrakeVEVO


Top 50 Trending Channel Titles in All Countries
ChildishGambinoVEVO 10652536179
NickyJamTV 9308934693
Ozuna 8478719125
Marvel Entertainment 7623099346
DrakeVEVO 7547934171
Bad Bunny 6989131765
ibighit 6483597160
ArianaGrandeVevo 6035313563
jypentertainment 5366841161
Ed Sheeran 5359375173
Flow La Movie 5168434018
TaylorSwiftVEVO 4701007199
YouTube Spotlight 4461592890
BeckyGVEVO 4372228601
MalumaVEVO 4162598454
Sony Pictures Entertainment 3636879404
Cardi B 3509689878
Disney•Pixar 3461395487
EminemVEVO 3330111174
Dude Perfect 3313743987
20th Century Fox 3311447602
jbalvinVEVO 3252008169
Kylie Jenner 3186921376
TheWeekndVEVO 3155692129
WORLDSTARHIPHOP 3028657807
Universal Pictures 2902075326
SebastianYatraVEVO 2850511592
MigosVEVO 2698494978
CalvinHarrisVEVO 2613451762
Maroon5VEVO 2553919603
PostMaloneVEVO 2518457440
Warner Bros. Pictures 2483367379
ZaynVEVO 2400641398
ChainsmokersVEVO 2332469930
NickiMinajAtVEVO 2320553848
Logan Paul Vlogs 2306144678
PewDiePie 2233358362
Daddy Yankee 2232028270
Pina Records 2110246118
EnriqueIglesiasVEVO 2066469248
ShawnMendesVEVO 1968000126
GEazyMusicVEVO 1960428194
KendrickLamarVEVO 1838109328
RomeoSantosVEVO 1831536664
FoxStarHindi 1823196576
SMTOWN 1694585148
MLG Highlights 1661126567
Bhad Bhabie 1645637816
JenniferLopezVEVO 1579451413
shakiraVEVO 1577147172
0.0
0.2
0.4
0.6
0.8
1.0
Views
1e10
Donyoe



Top Category ID