ASSIGNMENT 1: DATA LAKE WITH SNOWFLAKE

Simon Lim

24661225

Big Data Engineering

1 CONTEXTS AND OBJECTIVES OF THE PROJECT

YouTube is the worldwide video sharing platform, which share various categories and top trending videos on the platform. Top trending videos are evaluated based on video's number of views, shares, likes and comments etc. A dataset with a daily record of the top popular and trending videos has been extracted through YouTube API. In this regard, the objective of the project is to analyse the dataset, which consists of CSVs and Jsons files, by using a Data Lakehouse with Snowflake. The procedure of the project includes data ingestion, data cleaning, data analysis and answering business questions. Snowflake is a main data warehouse and SQL platform used in this project, in order to load and analyse data.

2 PRESENTATION OF DATASET

There are two types of dataset files, one with CSV file and the other with Json file. For CSV file, dataset includes several months of daily trending YouTube video records for 11 different countries, including India, USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea and Japan. Hence, there are 11 CSVs files representing top trending videos of each country. Various data of daily video records are included in dataset, such as video_id, title, published date, category id, number of views, likes, dislikes, channel id, comment counts and trending date etc.

For Json file, dataset also include category id, title and channel id for each country.

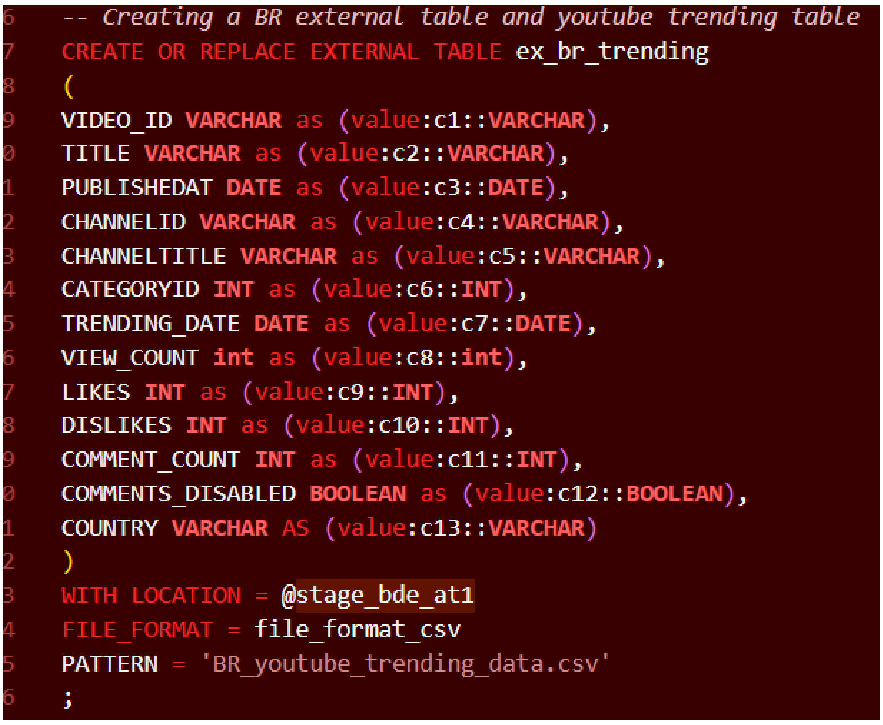
3 PIPELINE OF THE PROJECT WITH ISSUES

0. Before ingesting data, there are two necessary processes to be set up. The first step was to set up a cloud storage account on Microsoft Azure and snowflake account was a necessary process. Subsequently, it was also necessary to set up a Storage Integration between Azure and Snowflake.

1. Data Ingestion

1.1. Stage, named 'stage_bde_at1', was created to store, load and unload data files.

1.2. For CSVs files, external tables for each country were created, adding the new column 'country '. (11 External Tables). Columns in external tables include video_id, title, publish date, channel id, channel title, category id, trending date, view count, likes, dislikes, comment count, comment disabled and country.



```
-- Creating a BR external table and youtube trending table
CREATE OR REPLACE EXTERNAL TABLE ex_br_trending
(
VIDEO_ID VARCHAR as (value:c1::VARCHAR),
TITLE VARCHAR as (value:c2::VARCHAR),
PUBLISHEDAT DATE as (value:c3::DATE),
CHANNELID VARCHAR as (value:c4::VARCHAR),
CHANNELTITLE VARCHAR as (value:c5::VARCHAR),
CATEGORYID INT as (value:c6::INT),
TRENDING_DATE DATE as (value:c7::DATE),
VIEW_COUNT int as (value:c8::int),
LIKES INT as (value:c9::INT),
DISLIKES INT as (value:c10::INT),
COMMENT_COUNT INT as (value:c11::INT),
COMMENTS_DISABLED BOOLEAN as (value:c12::BOOLEAN),
COUNTRY VARCHAR AS (value:c13::VARCHAR)
)
WITH LOCATION = @stage_bde_at1
FILE_FORMAT = file_format_csv
PATTERN = 'BR_youtube_trending_data.csv'
;
```

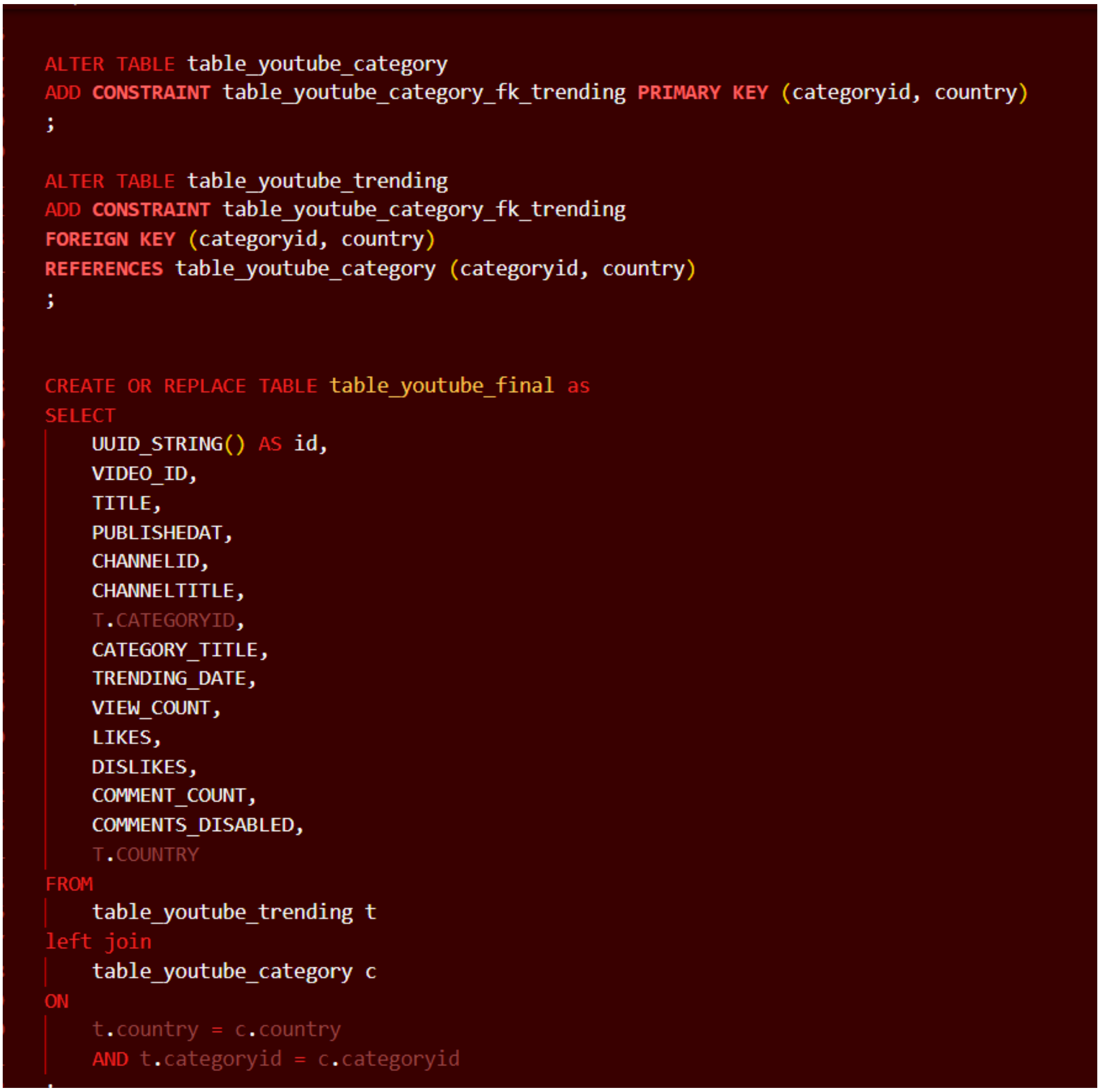
Figure 1. An External Table for Brazil

Then, a new table, ‘table_youtube_trending’ was created to store all data received from external tables. Subsequently, null values in the new column ‘country’ were updated after data for each country were inserted. External tables with same columns for other countries were all inserted into the table ‘table_youtube_trending’, in conjunction with updates of ‘country’ column.

1.3. For Json files, I manually added a country column and each corresponding country value. The rationale behind it was that it was possible to load 11 Json files as each row in an external table, helping to simplify code. Hence, an external table for 11 countries was created. Then, a new table ‘table_youtube_category’ was created to store data from the external table, including three columns, ‘country’, ‘category id’ and ‘category title’.

1.4. The last stage in data ingestion was to create a final table by combining two tables. Before combining them, it was necessary to set up primary keys and foreign keys (category id, country) for tables, as those two columns are common columns two tables included.

1.5. Finally, two tables were combined using left join on country and category id and created a final youtube table, “table_youtube_final”.



```
ALTER TABLE table_youtube_category
ADD CONSTRAINT table_youtube_category_fk_trending PRIMARY KEY (categoryid, country)
;

ALTER TABLE table_youtube_trending
ADD CONSTRAINT table_youtube_category_fk_trending
FOREIGN KEY (categoryid, country)
REFERENCES table_youtube_category (categoryid, country)
;


CREATE OR REPLACE TABLE table_youtube_final as
SELECT
    UUID_STRING() AS id,
    VIDEO_ID,
    TITLE,
    PUBLISHEDAT,
    CHANNELID,
    CHANNELTITLE,
    T.CATEGORYID,
    CATEGORY_TITLE,
    TRENDING_DATE,
    VIEW_COUNT,
    LIKES,
    DISLIKES,
    COMMENT_COUNT,
    COMMENTS_DISABLED,
    T.COUNTRY
FROM
    table_youtube_trending t
left join
    table_youtube_category c
ON
    t.country = c.country
    AND t.categoryid = c.categoryid
```

Figure 2. Combining two tables

2. Data Cleaning

In this stage, null values and duplicate values were cleaned in table_youtube_final table.

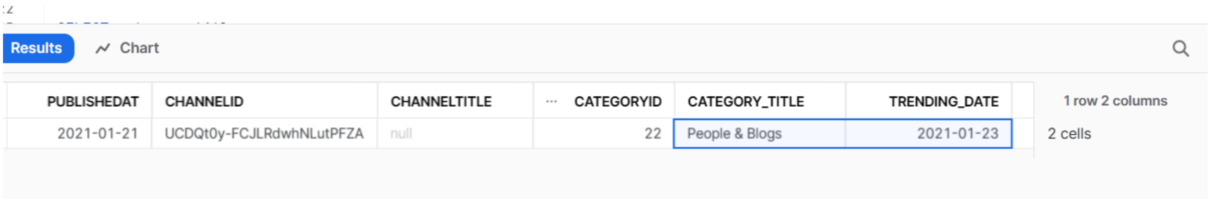
2.1. In “table_youtube_category”, only “Comedy” has duplicates if we do not take into account the categoryid.

2.2. In “table_youtube_category”, only “Nonprofit & Activism” appears in one country.

2.3. Categoryid, which has missing values in category_title, is ‘29’.

2.4. Null values (3161 rows) in category_title were filled with category id ‘29’.

2.5. There is one video, which has a null value for channeltitle.



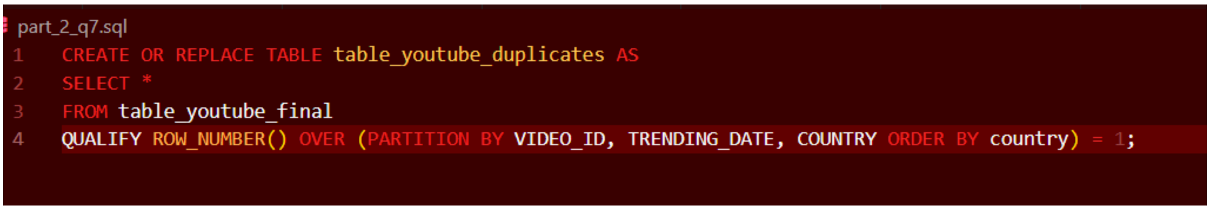
PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	CATEGORY_TITLE	TRENDING_DATE
2021-01-21	UCDQt0y-FCJLRdwhNLutPFZA	null	22	People & Blogs	2021-01-23

1 row 2 columns

2 cells

2.6. All records, containing ‘video_id’ of “#NAME?” were all deleted (15831 rows were deleted).

2.7. A duplicate table, “table_youtube_duplicates”, containing duplicates of a combination of video_id, trending_date and country, was created.



part_2_q7.sql

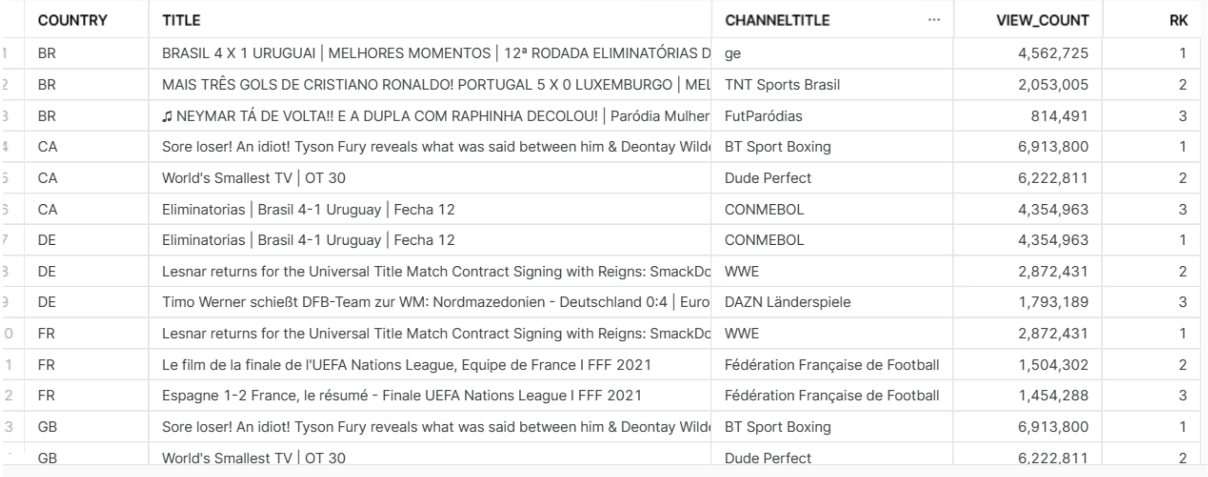
```sql
CREATE OR REPLACE TABLE table_youtube_duplicates AS
SELECT *
FROM table_youtube_final
QUALIFY ROW_NUMBER() OVER (PARTITION BY VIDEO_ID, TRENDING_DATE, COUNTRY ORDER BY country) = 1;
```

2.8. Finally, the duplicates in “table_youtube_final” were deleted using “table_youtube_duplicates”.

2.9. Final number of rows in table_youtube_final were 1123017.

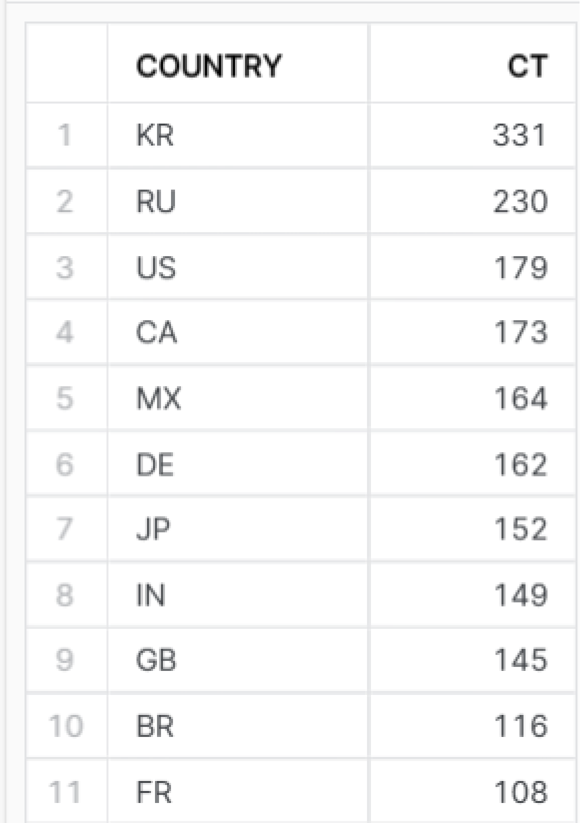
3. Data Analysis

3.1. In “table_youtube_final”, three most viewed videos in the “Sports” category and for the trending_date = “2021-10-17” for each country were analyzed and displayed.



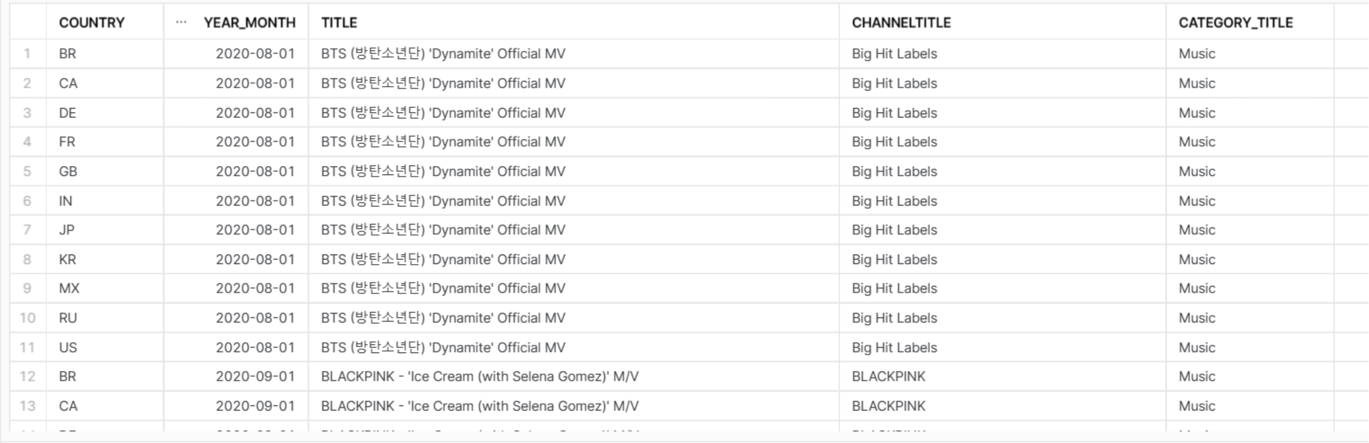
COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
BR	BRASIL 4 X 1 URUGUAI | MELHORES MOMENTOS | 12ª RODADA ELIMINATÓRIAS D	ge	4,562,725	1
BR	MAIS TRÊS GOLS DE CRISTIANO RONALDO! PORTUGAL 5 X 0 LUXEMBURGO | MEL	TNT Sports Brasil	2,053,005	2
BR	♫ NEYMAR TÁ DE VOLTA!! E A DUPLA COM RAPHINHA DECOLOU! | Paródia Mulher	FutParódias	814,491	3
CA	Sore loser! An idiot! Tyson Fury reveals what was said between him & Deontay Wild	BT Sport Boxing	6,913,800	1
CA	World's Smallest TV | OT 30	Dude Perfect	6,222,811	2
CA	Eliminatorias | Brasil 4-1 Uruguay | Fecha 12	CONMEBOL	4,354,963	3
DE	Eliminatorias | Brasil 4-1 Uruguay | Fecha 12	CONMEBOL	4,354,963	1
DE	Lesnar returns for the Universal Title Match Contract Signing with Reigns: SmackDc	WWE	2,872,431	2
DE	Timo Werner schießt DFB-Team zur WM: Nordmazedonien - Deutschland 0:4 | Euro	DAZN Länderspiele	1,793,189	3
FR	Lesnar returns for the Universal Title Match Contract Signing with Reigns: SmackDc	WWE	2,872,431	1
FR	Le film de la finale de l'UEFA Nations League, Equipe de France I FFF 2021	Fédération Française de Football	1,504,302	2
FR	Espagne 1-2 France, le résumé - Finale UEFA Nations League I FFF 2021	Fédération Française de Football	1,454,288	3
GB	Sore loser! An idiot! Tyson Fury reveals what was said between him & Deontay Wild	BT Sport Boxing	6,913,800	1
GB	World's Smallest TV | OT 30	Dude Perfect	6,222,811	2

3.2. The number of distinct video with a title containing the word “BTS” for each county was displayed.



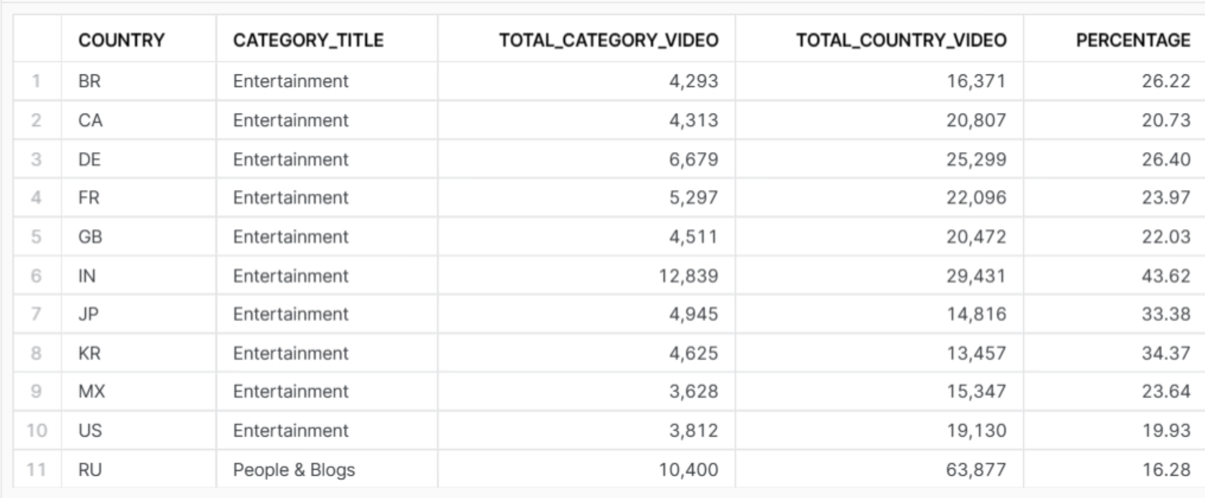
	COUNTRY	CT
1	KR	331
2	RU	230
3	US	179
4	CA	173
5	MX	164
6	DE	162
7	JP	152
8	IN	149
9	GB	145
10	BR	116
11	FR	108

3.3. The most viewed videos and their like_ratio were displayed with their year and month and for each country.



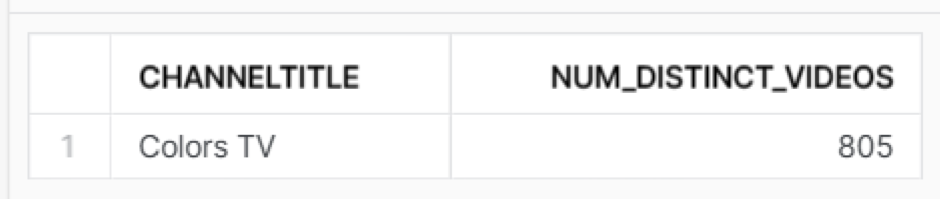
	COUNTRY	··· YEAR_MONTH	TITLE	CHANNELTITLE	CATEGORY_TITLE
1	BR	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
2	CA	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
3	DE	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
4	FR	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
5	GB	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
6	IN	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
7	JP	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
8	KR	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
9	MX	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
10	RU	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
11	US	2020-08-01	BTS (방탄소년단) 'Dynamite' Official MV	Big Hit Labels	Music
12	BR	2020-09-01	BLACKPINK - 'Ice Cream (with Selena Gomez)' M/V	BLACKPINK	Music
13	CA	2020-09-01	BLACKPINK - 'Ice Cream (with Selena Gomez)' M/V	BLACKPINK	Music

3.4. For each cointry, videos with a category_title, that has the most distinct videos and their percentage out of the total distinct number of videos of that country was displayed.



	COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
1	BR	Entertainment	4,293	16,371	26.22
2	CA	Entertainment	4,313	20,807	20.73
3	DE	Entertainment	6,679	25,299	26.40
4	FR	Entertainment	5,297	22,096	23.97
5	GB	Entertainment	4,511	20,472	22.03
6	IN	Entertainment	12,839	29,431	43.62
7	JP	Entertainment	4,945	14,816	33.38
8	KR	Entertainment	4,625	13,457	34.37
9	MX	Entertainment	3,628	15,347	23.64
10	US	Entertainment	3,812	19,130	19.93
11	RU	People & Blogs	10,400	63,877	16.28

3.5. Channeltitle, that has the most distinct videos and its number were displayed.

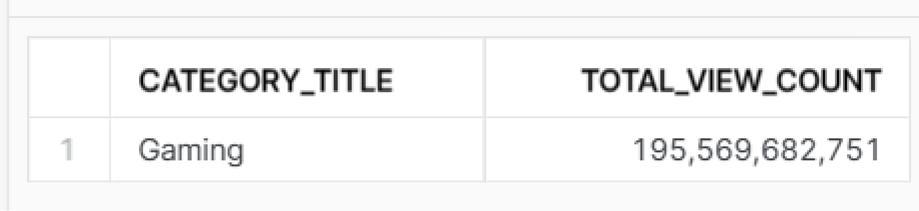


	CHANNELTITLE	NUM_DISTINCT_VIDEOS
1	Colors TV	805

4 ANSWERING BUSINESS QUESTIONS

1. If I was to launch a new YouTube channel, which category (excluding "Music" and "Entertainment") of video would I be trying to create to have them appear in the top trend of YouTube?

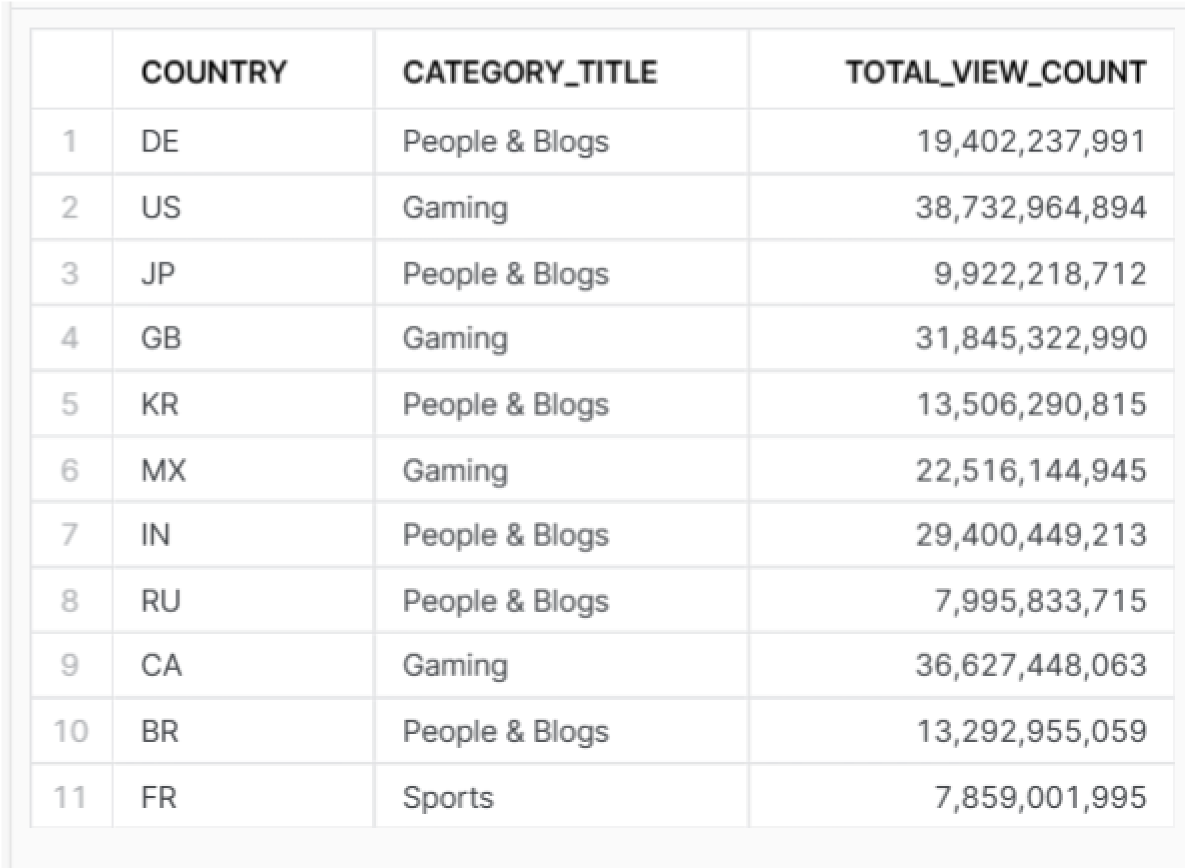
- I would create 'Gaming' category, as it has the highest total_view_count (195569682751) out of all categories.



	CATEGORY_TITLE	TOTAL_VIEW_COUNT
1	Gaming	195,569,682,751

2. Will this strategy work in every country?

- No, unfortunately it does not work in every country. There are many countries, in which "People & Blog" have higher total_view_count than "Gaming" category, including Japan, Korea, India, Russia, Brazil etc. Therefore, it is important to create a category of video, depending on country and other potential variables, such as age and gender.



	COUNTRY	CATEGORY_TITLE	TOTAL_VIEW_COUNT
1	DE	People & Blogs	19,402,237,991
2	US	Gaming	38,732,964,894
3	JP	People & Blogs	9,922,218,712
4	GB	Gaming	31,845,322,990
5	KR	People & Blogs	13,506,290,815
6	MX	Gaming	22,516,144,945
7	IN	People & Blogs	29,400,449,213
8	RU	People & Blogs	7,995,833,715
9	CA	Gaming	36,627,448,063
10	BR	People & Blogs	13,292,955,059
11	FR	Sports	7,859,001,995

5 REFERENCE

YouTube Trending Video Dataset (updated daily). (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset