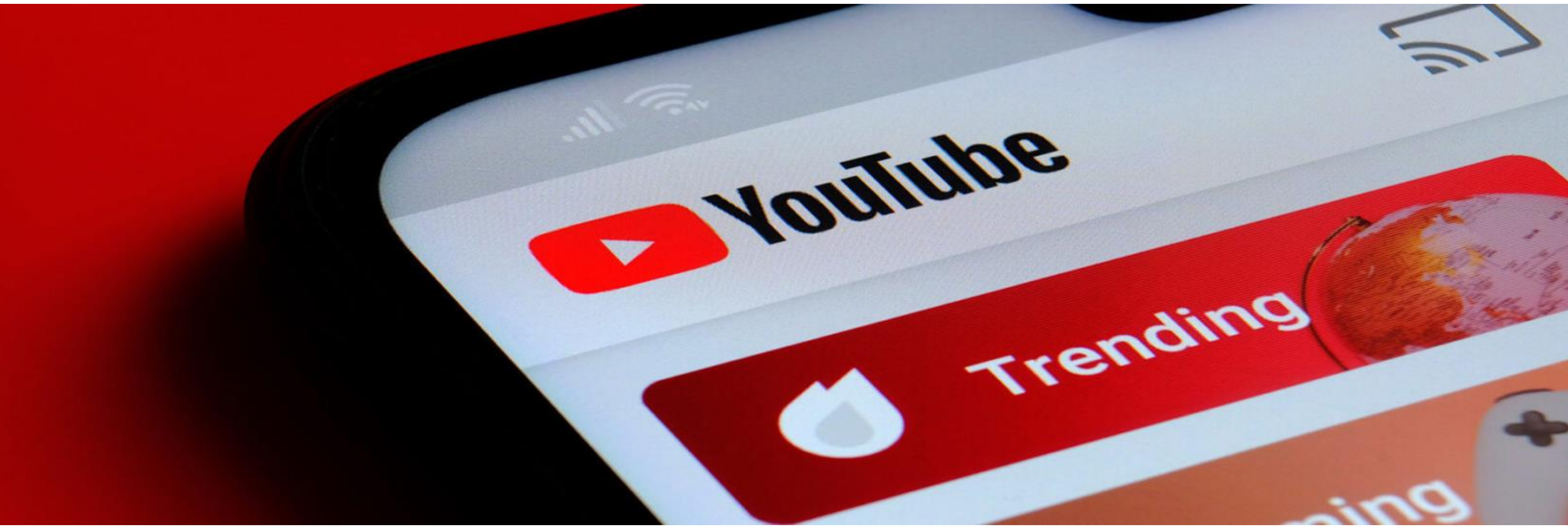
UNIVERSITY
OF TECHNOLOGY
SYDNEY

Determine YouTube trends with Snowflake



Big Data Engineering | Autumn, 2024

Taekjin Jeong (25099654)

<02.09.2024>

Table of Contents

1. Project Overview

This project utilises large volumes of daily updated YouTube data to build a data lakehouse, analyse the data, and gain insight. The first step is setting up Data Lakehouse using Snowflake through Azure (Content 2). In addition, this project involved data cleansing using SQL. Anomalies in the tables were identified, and unnecessary information was updated or deleted (Content 3). After that, we focus on Data analyses that provide an understanding of the dataset and have deepened perspectives (Content 4). Lastly, in business insight (Content 5), it shows the process of choosing the Science & technology category to launch a new business. Data analyses provide evidence of why this project chose one category. Every step has its own key points, and all queries will be provided with SQL files separately.

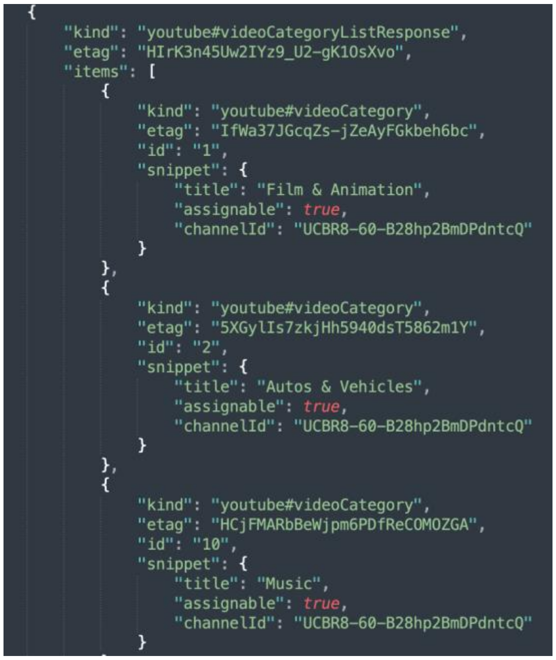
2. Data Lakehouse Setup & Data Ingestion

Data Lakeshouse provides both structured queries and advanced analytics are executed on the optimally structured data for the given usage (D. Oreščanin and T. Hlupić, 2021). This project develops Snowflake as a data lakehouse platform. Snowflake can run on multiple cloud platforms. It uses Azure as a Data infrastructure to upload the dataset to a storage. The dataset is given by Kaggle (https://www.kaggle.com/rsrishav/youtube-trending-video-dataset), and it is a daily record of top trending YouTube videos from 2020-08-12 to 2024-04-15). There are two datasets: the trending data consists of ten CSV files, while the category data is made up of ten JSON files. Each file includes ten countries data (Brazil, Canada, France, Germany, Great Britain, India, Japan, Korea, Mexico and the USA). Trending data includes the video title, channel title, published time, views, likes, dislikes and comments. Category data includes category ID and category title.

Table 1. Trending data sample

Columns	Data
video_id	s9FH4rDMvds
title	LEVEI UM FORA? FINGI ESTAR APAIXONADO POR ELA!
publishedAt	2020-08-11T22:21:49Z

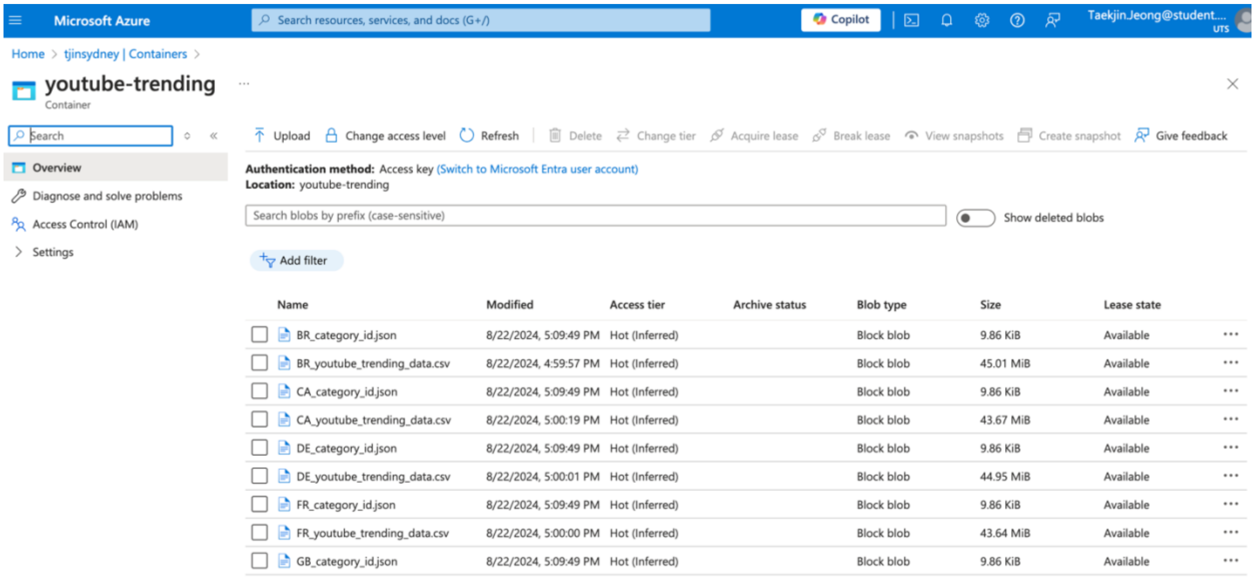
channelId	UCGfBwrCoi9ZJjKiUK8MmJNw
channelTitle	Pietro Guedes
categoryId	22
trending_date	2020-08-12T00:00:00Z
view_count	263835
likes	85095
dislikes	487
comment_count	4500



```
{
    "kind": "youtube#videoCategoryListResponse",
    "etag": "HIrK3n45Uw2IYz9_U2-gK10sXvo",
    "items": [
        {
            "kind": "youtube#videoCategory",
            "etag": "IfWa37JGcqZs-jZeAyFGkbeh6bc",
            "id": "1",
            "snippet": {
                "title": "Film & Animation",
                "assignable": true,
                "channelId": "UCBR8-60-B28hp2BmDPdntcQ"
            }
        },
        {
            "kind": "youtube#videoCategory",
            "etag": "5XGylIs7zkjHh5940dsT5862m1Y",
            "id": "2",
            "snippet": {
                "title": "Autos & Vehicles",
                "assignable": true,
                "channelId": "UCBR8-60-B28hp2BmDPdntcQ"
            }
        },
        {
            "kind": "youtube#videoCategory",
            "etag": "HCjFMARbBeWjpm6PDfReCOMOZGA",
            "id": "10",
            "snippet": {
                "title": "Music",
                "assignable": true,
                "channelId": "UCBR8-60-B28hp2BmDPdntcQ"
            }
```

Figure 1. Category data sample

The first step in this part involves uploading the dataset to Azure for data storage. A new container named ‘YouTube-trending’ is created. To access this container through Snowflake, a SAS token is generated and used in a query to connect the storage named stage_assignment.



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: youtube-trending

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
□	BR_category_id.json	8/22/2024, 5:09:49 PM	Hot (Inferred)		Block blob	9.86 KiB	Available	···
□	BR_youtube_trending_data.csv	8/22/2024, 4:59:57 PM	Hot (Inferred)		Block blob	45.01 MiB	Available	···
□	CA_category_id.json	8/22/2024, 5:09:49 PM	Hot (Inferred)		Block blob	9.86 KiB	Available	···
□	CA_youtube_trending_data.csv	8/22/2024, 5:00:19 PM	Hot (Inferred)		Block blob	43.67 MiB	Available	···
□	DE_category_id.json	8/22/2024, 5:09:49 PM	Hot (Inferred)		Block blob	9.86 KiB	Available	···
□	DE_youtube_trending_data.csv	8/22/2024, 5:00:01 PM	Hot (Inferred)		Block blob	44.95 MiB	Available	···
□	FR_category_id.json	8/22/2024, 5:09:49 PM	Hot (Inferred)		Block blob	9.86 KiB	Available	···
□	FR_youtube_trending_data.csv	8/22/2024, 5:00:00 PM	Hot (Inferred)		Block blob	43.64 MiB	Available	···
□	GB_category_id.json	8/22/2024, 5:09:49 PM	Hot (Inferred)		Block blob	9.86 KiB	Available	···

Figure 1. Upload files on Azure

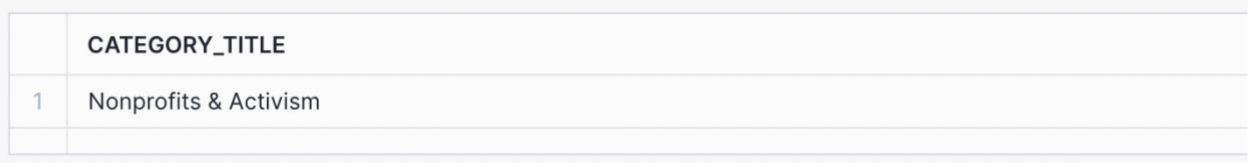
Executing the command LIST @stage_assignment on Snowflake verifies the connection between Snowflake and Azure. To create the inner table, we set up external tables first. Snowflake provides functions when creating tables and the split_part function was used to extract country data from the file names. table_youtube_trending and table_youtube_category were created from external tables. After that, the id column was added by using the UUID_STRING() function to add a disguisable identifier column. It returns a 128-bit value, formatted as a string (Snowflake documentation). After all data ingestion, The Final Table (table_youtube_final) has 2,667,041 rows.

Key Points

- Snowflake provides useful documentation about functions. It shares sample queries to help understand functions. (https://docs.snowflake.com/en/sql-reference/functions/)
- When creating the **table_Youtube_final,** the country information was sourced from the category table. Although the total numbers initially matched the target results, discrepancies were identified after data cleansing, revealing that the results differed from the expected outcome. Therefore, it is crucial to thoroughly verify the source of table columns when initially designing the table.

3. Data Cleaning

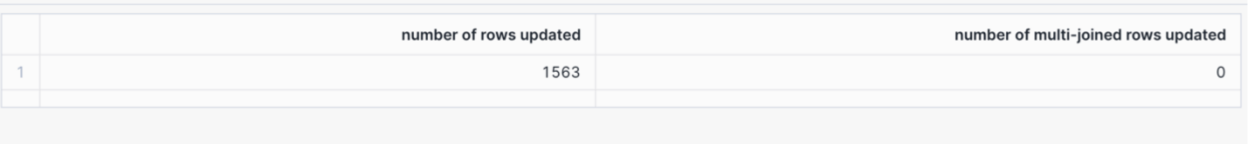
In table_youtubue_category, it has 31 features and a duplication of category_title. A single country should have only one category, but 'Comedy' was associated with more than one category. Furthermore, **‘Nonprofits & Activism’** only appeared in one country.



	CATEGORY_TITLE
1	Nonprofits & Activism

Figure 2. Category title that appeared in one country

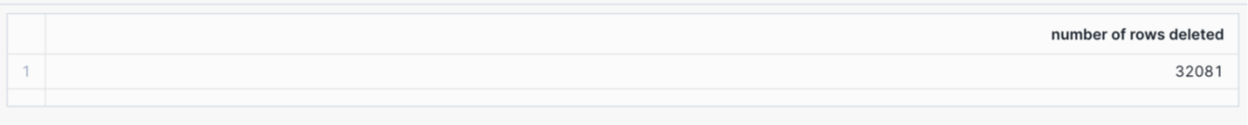
This project focuses on data analysis as the next step. Data cleaning was essential to provide premise results. The final table has missing category titles where the category ID is **‘29’**, hence 1,563 rows are updated in the final table by the category table.



	number of rows updated	number of multi-joined rows updated
1	1563	0

Figure 3. The result of updating the final table

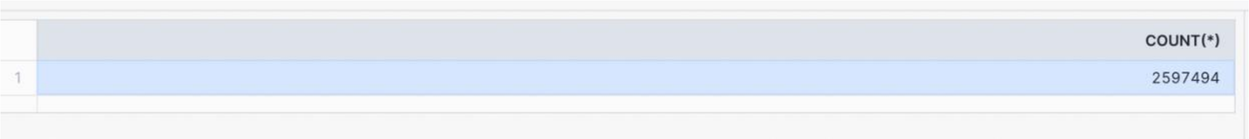
After that, rows where the video ID is “#name?” were deleted to make sure all video IDs have distinct values. **32,081 rows** are removed.



	number of rows deleted
1	32081

Figure 4. The result of deleting unnamed rows

The final table contains duplicate rows even if they have the same video ID, country, and trending date. However, they have different metrics, such as likes and dislikes. We created another table called **‘table_youtube_duplicates’** to define duplicated rows. When creating a new table, **ROW_NUMBER()** function is used. It Returns a unique row number for each row within a window partition (Snowflake documentation). This final table has 37,466 duplicated rows. We removed those rows from the final table, and 2,597,494 rows remained.



	COUNT(*)
1	2597494

Figure 5. The result after Data cleansing

Key Points

- When creating the **table_Youtube_duplicates,** it is vital to ensure the partitioning by country, video ID and trending date. If we did not partition these columns, the result will be changed.
- When deleting data compared to **table_Youtube_duplicates**, it is essential to include `id`, `video_id`, `country`, and `trending_date` in the WHERE clause. Failing to do so may result in deleting data that does not match the number in the duplicate table."

4. Data Analyses

- Q1) The three most viewed videos for each country in the Gaming category on 4 April 2024

- Table 3 *(see appendices)* shows the result of Question 1. It indicates that “DAGGER DUCHESS – New Tower Troop!” is the most famous video in Gaming among countries. However, in Asia, there are different videos compared to other countries.

- Q2) For each country, count the number of distinct videos with a title containing the word “BTS”

- Table 4 *(see appendices)* shows that South Korea has the highest number of videos related to BTS Since BTS is a South Korean band. The query includes a subquery and LOWER function to distinguish videos with the word "bts". **LOWER** function Returns the input string with all characters converted to lowercase.

- Q3) For each country, year and month (in a single column) and only for the year for 2024, which video is the most viewed and what is its likes_ratio (defined as the percentage of likes against view_count) truncated to 2 decimals.

  - Table 5 *(see appendices)* is a Pivot table of Q3 query results. It indicates that in most countries, the most viewed videos are Entertainment videos, and most of them are from the channel called 'MrBeast". The subquery includes RANK() and EXTRACT functions to get data for 2024. Furthermore, TRUNCATE function helps to round truncate to 2 decimals. It Rounds the input expression down to the nearest (or equal) integer closer to zero or the nearest equal or smaller value with the specified number of places after the decimal point. (Snowflake documentation)

- Q4) For each country, which category title has the most distinct videos and what is its percentage (2 decimals) out of the total distinct number of videos of that country from 2022.
  - Table 6 *(see appendices)* is the result of question 4. It shows that the most distinct videos are about Entertainment in all the countries except two countries (Canada and the US). Entertainment is a significant category in India, accounting for over 42% of all videos. To get the result, this query includes CTE. A Common Table Expression (CTE) is a temporary result set that you can define within an SQL query. CTE counts total category videos and ranks them by country. total_country_video column sums all categories of videos counted by countries.

- Q5) Which *channeltitle* has produced the most **distinct** videos, and what is this number
  - The channel **'Vijay Television'** creates the most distinct videos and counting 2,049. The query counts videos by channel title and uses Limit 1 to show the highest distinct video channel.

Key Points

- When a subquery is long, it can make a query more challenging to read and understand. However, by using a CTE, we can improve readability and create a query that is easier for others to understand.

5. Business Insight

By understanding the YouTube dataset, this project aims to choose one category to launch new channel for a new business. We excluded Entertainment and Music category. To begin with, we made a query to see categories change by year. Figure 7 shows the top five categories. Among categories, Gaming is the highest category people viewed. However, Gaming has 2,622 channels in 2023, so if we divide view counts by channels, then it marks a lower ratio in Gaming. It implies Gaming category is very competitive even though it is a top category. Moreover, Films & Animation experienced the highest growth from 2022 to 2023, with a 45% increase, and Science & Technology increased by 42%. They are also less competitive within the top five categories.

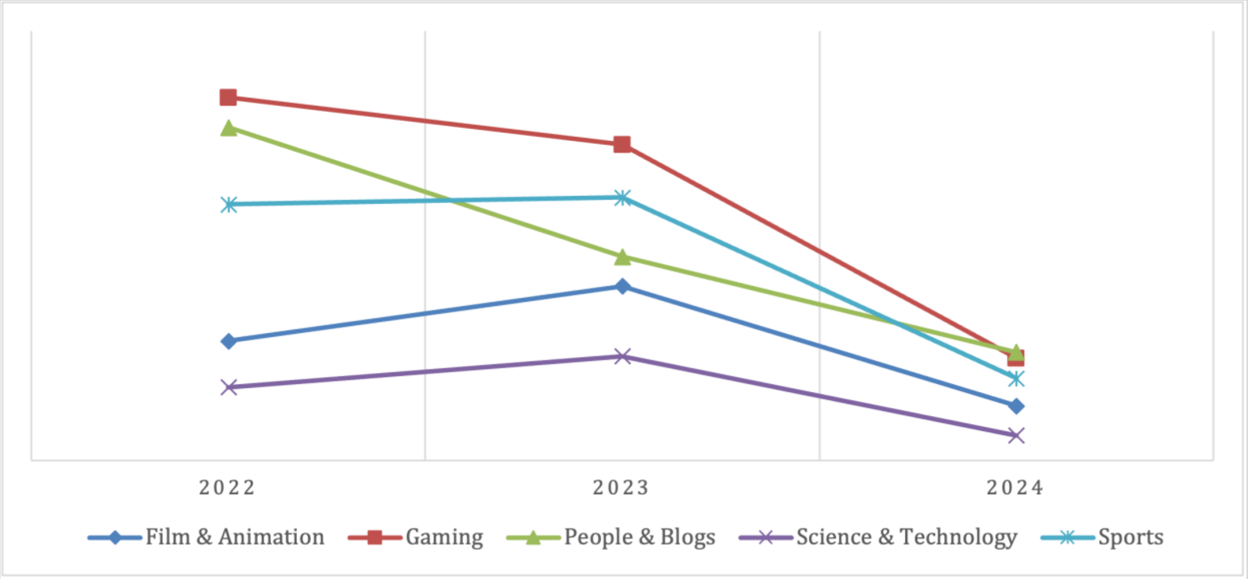


Figure 6. Top five categories

Based on the above analyses, we have narrowed down the category for our launch to **Films & Animation and Science & Technology**. By analysing the top five videos by country, we can choose one category for a new business. Table 2 shows aggregated results of the top five videos by categories after 2023 and highlights that **Science & Technology** has 213 videos the most**.** Total view counts also double that of Film & Animation.

Table 2. The aggregated results of the top 5 videos by categories after 2023

CATEGORY_TITLE	CATEGORY_COUNT	TOTAL_VIEW_COUNT	CHANNEL_COUNT	VIDEO_COUNT
Science & Technology	213	4885261844	16	46
Film & Animation	144	2114047244	20	60
Sports	138	2351385878	23	50
Gaming	132	2508141144	22	43
People & Blogs	84	949789563	28	56
Comedy	43	509599845	14	27
News & Politics	21	238547553	9	13

Howto & Style	12	528544393	3	4
Autos & Vehicles	7	76864490	3	3
Pets & Animals	2	14654036	2	2
Education	2	19068650	2	2
Nonprofits & Activism	1	5391630	1	1
Travel & Events	1	3863077	1	1

To confirm that **Science & Technology** category is applicable in every country, we compare five categories by country. Figure 8 confirms that in seven countries, excluding Japan, South Korea, and Mexico, **Science & Technology** ranks either first or second.

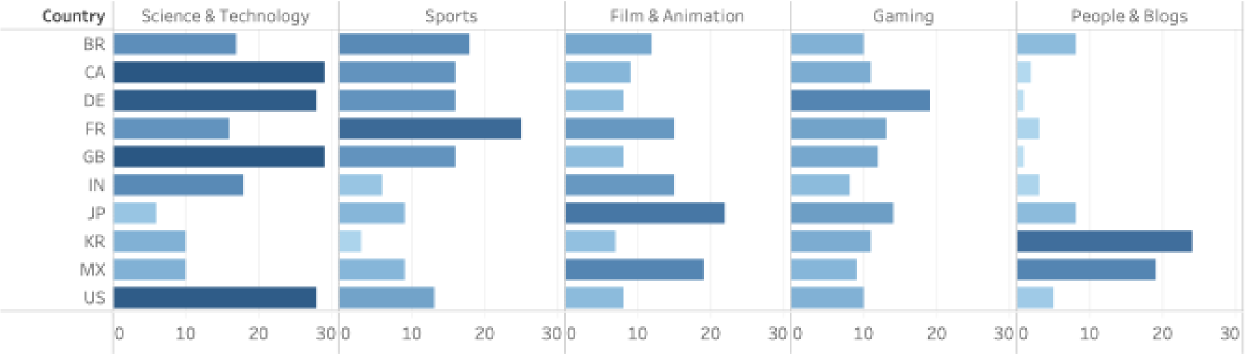


Figure 7. The comparison of categories by countries

As a result, our project chooses **Science & Technology** to launch a new business. **Science & Technology** is less competitive than other categories, and the likelihood of ranking in the top five is higher. Given that achieving a top five rank would have a global impact, we have determined that this category is suitable for a new business.

Key Points

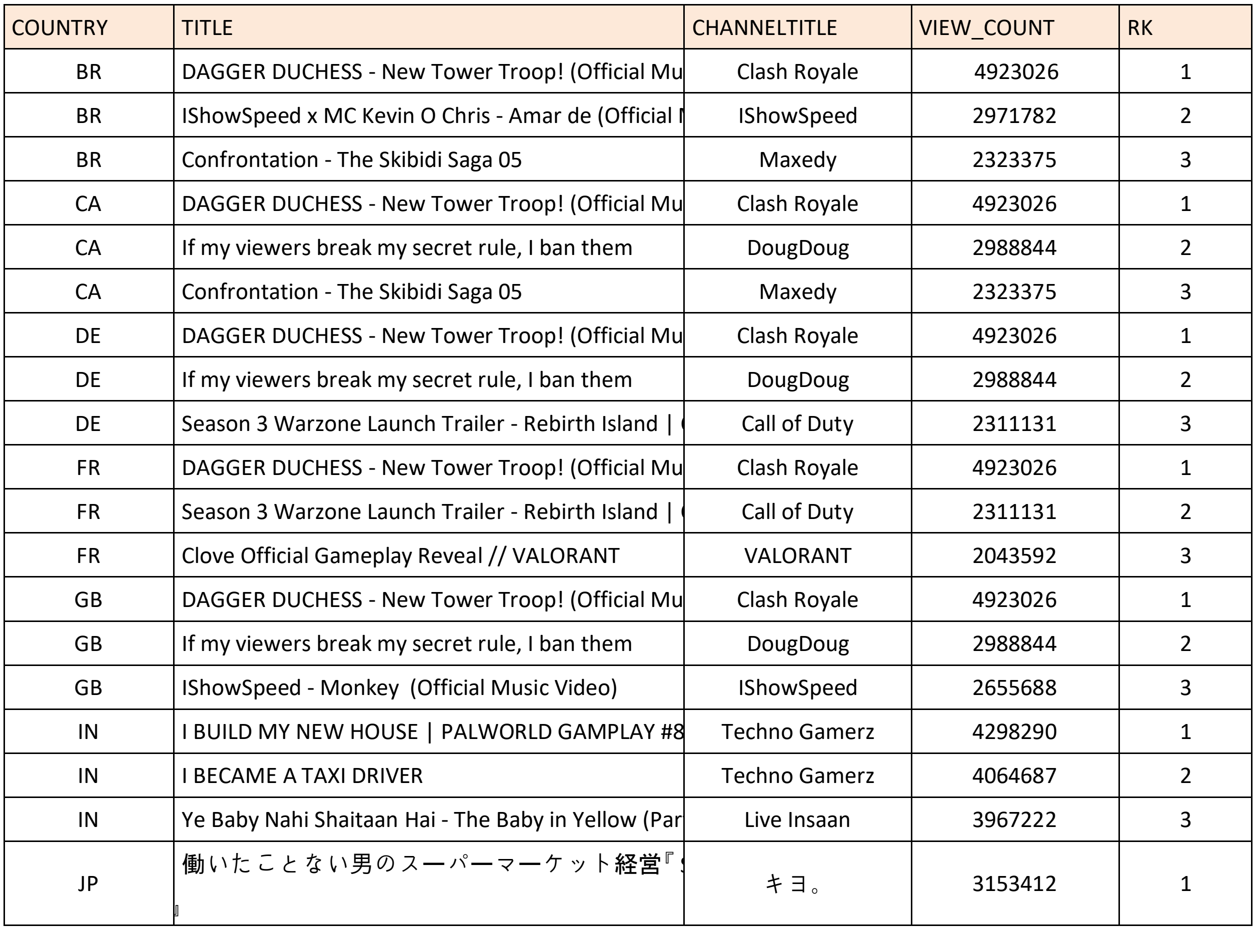
- Although Entertainment and Music categories were excluded, it was observed that many videos in the People & Blogs and Gaming categories also had an entertainment aspect. Therefore, several conditions were added to exclude these videos from the classification.

5. References

- SnowFlake documentation (https://docs.snowflake.com/)
- D. Oreščanin and T. Hlupić, "Data Lakehouse - a Novel Step in Analytics Architecture," 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2021, pp. 1242-1246, doi: 10.23919/MIPRO52101.2021.9597091.

6. Appendices

Table 3. Data Analysis Q1 result



COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
BR	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	1
BR	IShowSpeed x MC Kevin O Chris - Amar de (Official	IShowSpeed	2971782	2
BR	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
CA	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	1
CA	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
CA	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
DE	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	1
DE	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
DE	Season 3 Warzone Launch Trailer - Rebirth Island |	Call of Duty	2311131	3
FR	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	1
FR	Season 3 Warzone Launch Trailer - Rebirth Island |	Call of Duty	2311131	2
FR	Clove Official Gameplay Reveal // VALORANT	VALORANT	2043592	3
GB	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	1
GB	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
GB	IShowSpeed - Monkey (Official Music Video)	IShowSpeed	2655688	3
IN	I BUILD MY NEW HOUSE | PALWORLD GAMPLAY #8	Techno Gamerz	4298290	1
IN	I BECAME A TAXI DRIVER	Techno Gamerz	4064687	2
IN	Ye Baby Nahi Shaitaan Hai - The Baby in Yellow (Par	Live Insaan	3967222	3
JP	働いたことない男のスーパーマーケット経営『 』	キヨ。	3153412	1

JP	働いたことない男がバイトを雇うスーパーマー ator 』	キヨ。	2116115	2
JP	【崩壊：スターレイル】黄泉 キャラクターPV「君	崩壊：スターレイル	1206271	3
KR	'HOT DEBUT' 아일릿(ILLIT) - Magnetic #엠카운트ㄷ 송	Mnet K-POP	1678685	1
KR	(여자)아이들 - 나는 아픈 건 딱 질색이니까 #엠카 21 방송	Mnet K-POP	1337298	2
KR	Kissing You(키싱유) 이세계아이돌 COVER ｜[소녀 유)] (ISEGYE IDOL)	왁타버스 WAKTAVERSE	1093916	3
MX	HAPPY WHEELS, PERO EN 2024 !! - Episodio 28 | Fe	Fernanfloo	7227426	1
MX	ENTRE AL TUNEL DE ALFA EN ROBLOX!	FedeGames	5938537	2
MX	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	3
US	DAGGER DUCHESS - New Tower Troop! (Official Mu	Clash Royale	4923026	1
US	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
US	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3

Table 4. Data Analysis Q2 result

COUNTRY	CT
KR	494
IN	294
US	271
CA	264
MX	258
JP	257
DE	246
GB	227
BR	189
FR	172

Table 5. Data Analysis Q3 result

CHANNEL TITLE	Column Labels				
	Jan	Feb	Mar	Apr	Total
Entertainment	**5**	**10**	**10**	**8**	**33**
BR	1	1	1	1	4
CA		1	1	1	3
DE		1	1	1	3
FR		1	1	1	3

GB		1	1	1	3
IN	1	1	1	1	4
JP	1	1	1		3
KR	1	1	1		3
MX	1	1	1	1	4
US		1	1	1	3
Gaming	**5**				**5**
CA	1				1
DE	1				1
FR	1				1
GB	1				1
US	1				1
People & Blogs				**2**	**2**
JP				1	1
KR				1	1

Table 6. Data Analysis Q4 result

COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
BR	Entertainment	5417	23769	22.79
DE	Entertainment	7709	30759	25.06
FR	Entertainment	7548	32866	22.96
GB	Entertainment	5643	27873	20.24
IN	Entertainment	21281	50280	42.32
JP	Entertainment	5658	17645	32.06
KR	Entertainment	5122	15196	33.7
MX	Entertainment	4195	17545	23.9
CA	Gaming	6594	30886	21.34
US	Gaming	6226	28817	21.6