Big Data Engineering

Assignment 1

Ronik Jayakumar
24680264
ronik.jayakumar@student.uts.edu.au

1. Project Overview:

YouTube, with over 2.7 billion monthly active users, is the second most popular social media platform globally, behind Facebook. It reaches 52% of the world's 5.17 billion social media users. The platform's popularity has made it a top choice for content creators, who can join YouTube's partner program after gaining 1,000 subscribers and 4,000 watch hours. This allows them to monetise through ads, fan funding, and merchandise sales. In 2023, YouTube's top earner, Mr. Beast, made $82 million, highlighting the platform's lucrative potential.

Data related to trending videos and their categories for 10 countries has been obtained and analysed to understand trends across views, likes, dislikes, comments, etc. across the platform for the videos posted in these countries. Using these insights, a business plan is being constructed on the category with the highest overall potential to top the trending charts. This business plan is being constructed ignoring the music and entertainment industry.

2. Database Setup:

2.1 Microsoft Azure:

Microsoft azure was used as the database storage space for this project. The steps taken on Azure were as follows:

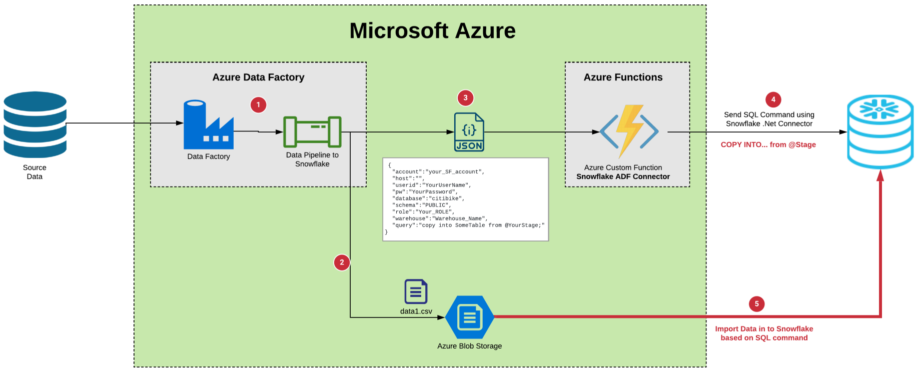
- Setup a storage account
- Add a storage container
- Retrieve the SAS token and the URL

2.2 Snowflake:

Snowflake has been considered as the platform of choice to run the SQL queries on our database. The reasoning for this is the high processing speed and the inbuilt availability of SQL within snowflake.

The steps taken within snowflake are as follows:

- Create a new database (which in our case has been called assignment_1)
- Create a new stage that points towards our storage container in azure using the retrieved URL and SAS token.
- Import the tables that have been uploaded onto Azure into our Snowflake environment as external tables.



```
{
 "account":"your_SF_account",
 "host":"",
 "userid":"YourUserName",
 "pw":"YourPassword",
 "database":"citibike",
 "schema":"PUBLIC",
 "role":"Your_ROLE",
 "warehouse":"Warehouse_Name",
 "query":"copy into SomeTable from @YourStage;"
}
```

3. Dataset & Preprocessing:

The dataset used was the trending videos YouTube data for 10 countries namely - Brazil, Canada, Denmark, France, Great Britain, India, Japan, Korea, Mexico, and the US. Along with this, the category data for these videos were also obtained for the above countries.

The two datasets were imported into snowflake as external tables and preprocessing steps were carried out on them. The overall preprocessing can be split into two major portions, External and internal tables.

3.1 External Tables section:

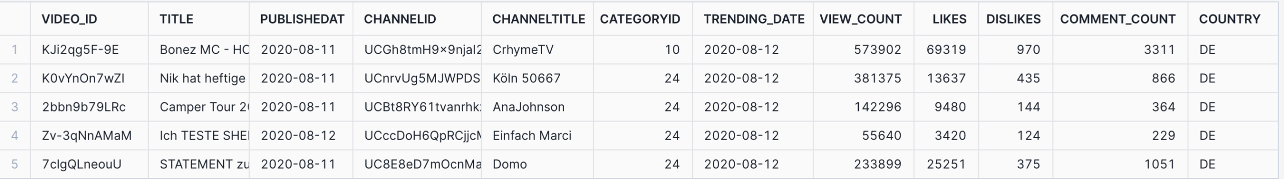
- **Standard File Format for CSV Files**:
  - A unified file format was created for all trending video files in CSV format. This ensured consistent handling of null values and the consolidation of all trending video files into a single table.
  - Action: Imported the trending video data into a consolidated table named *ex_table_youtube_trending*.

- **Standard File Format for JSON Files:**
  - A similar approach was applied to category information files in JSON format, allowing all category data to be merged seamlessly.
  - Action: Imported all category-related data into a table named *ex_table_youtube_category*.

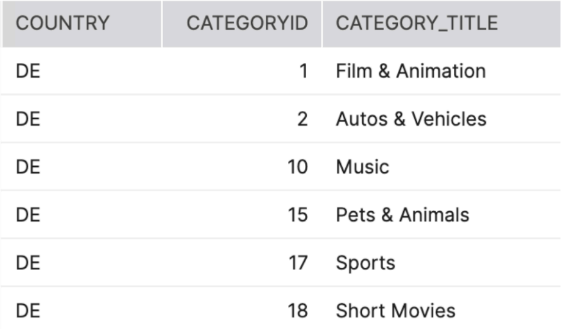
- **Data Type Standardisation:**
  - Ensured all columns, such as PublishedAt, were of the correct datatype (e.g., date) across the tables.

- **Country Column Parsing:**
  - Extracted the country information from the file names and added it as a Country column in the *ex_table_youtube_trending* table.

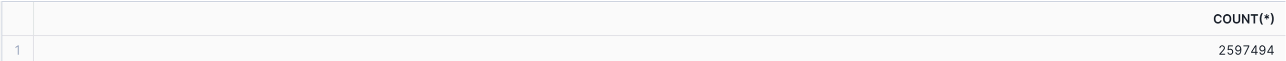
A snippet of both the tables can be seen below:



	VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	TRENDING_DATE	VIEW_COUNT	LIKES	DISLIKES	COMMENT_COUNT	COUNTRY
1	KJi2qg5F-9E	Bonez MC - HC	2020-08-11	UCGh8tmH9×9njal2	CrhymeTV	10	2020-08-12	573902	69319	970	3311	DE
2	K0vYnOn7wZI	Nik hat heftige	2020-08-11	UCnrvUg5MJWPDS	Köln 50667	24	2020-08-12	381375	13637	435	866	DE
3	2bbn9b79LRc	Camper Tour 2(	2020-08-11	UCBt8RY61tvanrhk:	AnaJohnson	24	2020-08-12	142296	9480	144	364	DE
4	Zv-3qNnAMaM	Ich TESTE SHEI	2020-08-12	UCccDoH6QpRCjjcN	Einfach Marci	24	2020-08-12	55640	3420	124	229	DE
5	7clgQLneouU	STATEMENT zu	2020-08-11	UC8E8eD7mOcnMa	Domo	24	2020-08-12	233899	25251	375	1051	DE



COUNTRY	CATEGORYID	CATEGORY_TITLE
DE	1	Film & Animation
DE	2	Autos & Vehicles
DE	10	Music
DE	15	Pets & Animals
DE	17	Sports
DE	18	Short Movies



	COUNT(*)
1	2597494

3.2 Internal Table & Preprocessing

- **Combining External Tables:**
  - An internal table named table_youtube_final was created by combining records from *ex_table_youtube_trending* and *ex_table_youtube_category*.

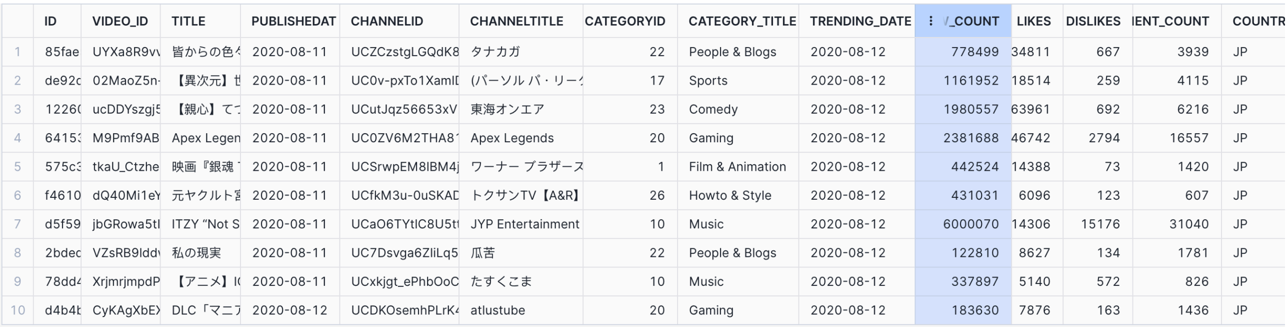
- **Handling Null Values:**
  - Replaced null values in critical fields such as `category_title` and `categoryid` to ensure data integrity.

- **Title Column Cleanup:**
  - Removed records with empty or missing titles to maintain data quality.

- **Duplicate Handling:**
  - Identified and removed duplicate records based on the title column, retaining only the record with the highest view_count.

- **Unique Identifier Addition:**
  - Introduced a unique identifier column using the UUID_STRING() function to maintain uniqueness across records.

The final table contains **2,597,494** rows and **14** columns. A snippet has been attached below.



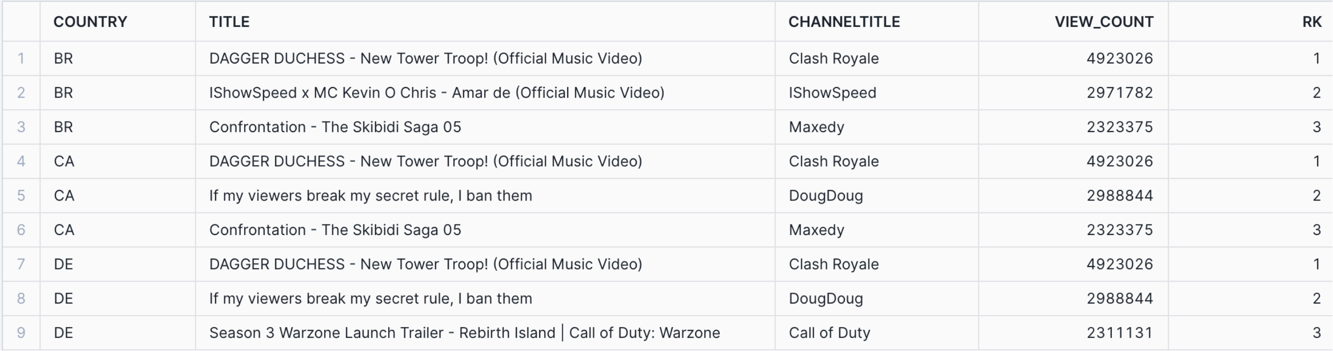
	ID	VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	CATEGORY_TITLE	TRENDING_DATE	⋮ /_COUNT	LIKES	DISLIKES	IENT_COUNT	COUNTR
1	85fae	UYXa8R9vv	皆からの色々	2020-08-11	UCZCzstgLGQdK8	タナカガ	22	People & Blogs	2020-08-12	778499	34811	667	3939	JP
2	de92c	02MaoZ5n-	【異次元】せ	2020-08-11	UC0v-pxTo1XamlD	(パーソル パ・リーグ	17	Sports	2020-08-12	1161952	18514	259	4115	JP
3	1226(	ucDDYszgj5	【親心】てつ	2020-08-11	UCutJqz56653xV	東海オンエア	23	Comedy	2020-08-12	1980557	63961	692	6216	JP
4	6415:	M9Pmf9AB	Apex Legen	2020-08-11	UC0ZV6M2THA8	Apex Legends	20	Gaming	2020-08-12	2381688	46742	2794	16557	JP
5	575c3	tkaU_Ctzhe	映画『銀魂	2020-08-11	UCSrwpEM8lBM4j	ワーナー ブラザース	1	Film & Animation	2020-08-12	442524	14388	73	1420	JP
6	f4610	dQ40Mi1eY	元ヤクルト宮	2020-08-11	UCfkM3u-0uSKAD	トクサンTV【A&R】	26	Howto & Style	2020-08-12	431031	6096	123	607	JP
7	d5f59	jbGRowa5tl	ITZY “Not S	2020-08-11	UCaO6TYtlC8U5tt	JYP Entertainment	10	Music	2020-08-12	6000070	14306	15176	31040	JP
8	2bdec	VZsRB9ldd	私の現実	2020-08-11	UC7Dsvga6ZliLq5	瓜苦	22	People & Blogs	2020-08-12	122810	8627	134	1781	JP
9	78dd4	XrjmrjmpdP	【アニメ】I(	2020-08-11	UCxkjgt_ePhbOoC	たすくこま	10	Music	2020-08-12	337897	5140	572	826	JP
10	d4b4b	CyKAgXbEX	DLC「マニア	2020-08-12	UCDKOsemhPLrK4	atlustube	20	Gaming	2020-08-12	183630	7876	163	1436	JP

4. Data Analysis:

The data analysis part faced us with 5 key questions on the database for which some key SQL tools have been utilised.

Question 1: What are the 3 most viewed videos for each country in the Gaming Category for the trending date = “2024-04-01”, Order the result by country and rank

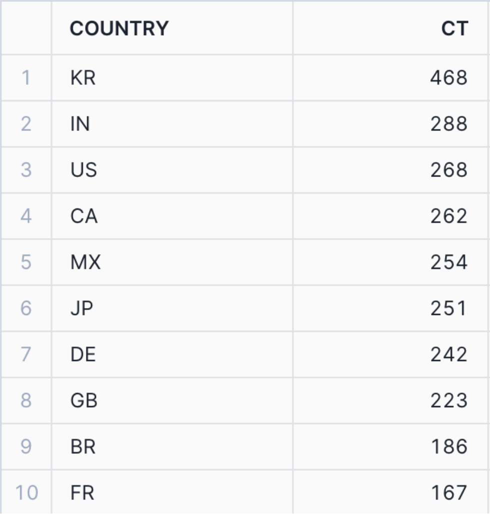
This query resulted in a table comprising of 30 columns with the top 3 videos for each country in the “Gaming” category listed out. For the trending date "2024-04-01" in the Gaming category, the most viewed video across multiple countries was "DAGGER DUCHESS - New Tower Troop! (Official Music Video)" by Clash Royale. It ranked first in several countries, including Brazil, Canada, Germany, France, the UK, Mexico, and the US. Other notable videos, such as "If my viewers break my secret rule, I ban them" by DougDoug and various other gaming-related videos, also featured prominently across these countries in the top 3. A snippet of the result is attached below.



	COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
1	BR	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
2	BR	IShowSpeed x MC Kevin O Chris - Amar de (Official Music Video)	IShowSpeed	2971782	2
3	BR	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
4	CA	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
5	CA	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
6	CA	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
7	DE	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
8	DE	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
9	DE	Season 3 Warzone Launch Trailer - Rebirth Island | Call of Duty: Warzone	Call of Duty	2311131	3

Question 2: For each country, count the number of distinct video with a title containing the word “BTS” (case insensitive) and order the result by count in a descending order.

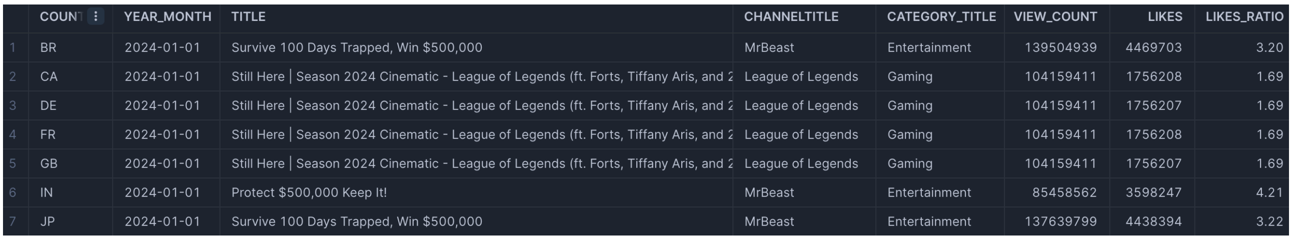
The query results in a table comprising of 10 rows, 1 for each country corresponding with its overall number of occurrences with South Korea (KR) having the highest count at 468. India (IN) and the United States (US) followed with 288 and 268 and with France (FR) having the lowest count at 167.



	COUNTRY	CT
1	KR	468
2	IN	288
3	US	268
4	CA	262
5	MX	254
6	JP	251
7	DE	242
8	GB	223
9	BR	186
10	FR	167

Question 3: For each country, year and month and only for the year 2024, which video is the most viewed and what is its likes_ratio truncated to 2 decimals. Order the result by year_month and country.

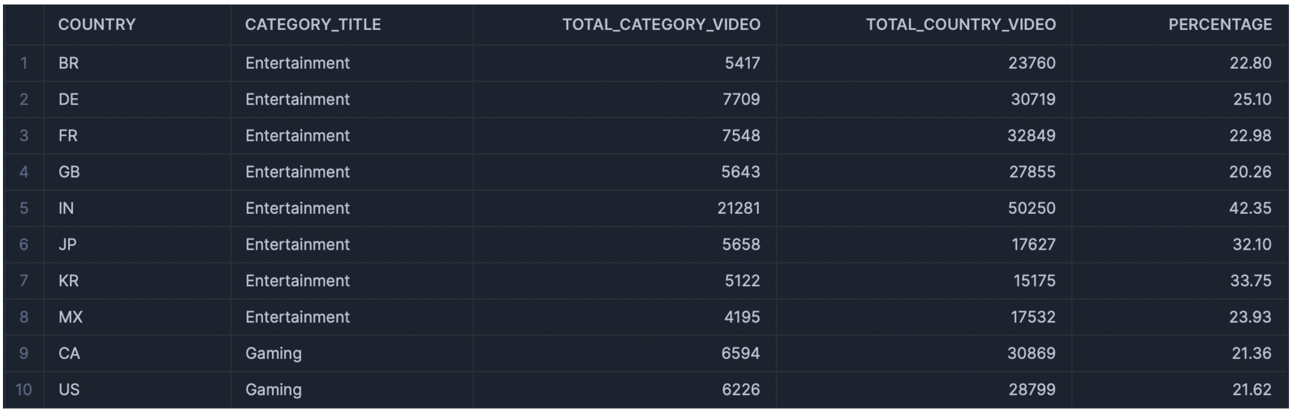
In this question, a deeper level of analysis has been carried out to identify the best performing video in terms of highest views for each country for each month of 2024. Using the view count and the overall likes that video received, the likes_ratio has also been calculated. The dates have been truncated to show the first of each month in every occurrence. The final results obtained were as follows:



	COUN	YEAR_MONTH	TITLE	CHANNELTITLE	CATEGORY_TITLE	VIEW_COUNT	LIKES	LIKES_RATIO
1	BR	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	139504939	4469703	3.20
2	CA	2024-01-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2	League of Legends	Gaming	104159411	1756208	1.69
3	DE	2024-01-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2	League of Legends	Gaming	104159411	1756207	1.69
4	FR	2024-01-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2	League of Legends	Gaming	104159411	1756208	1.69
5	GB	2024-01-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2	League of Legends	Gaming	104159411	1756207	1.69
6	IN	2024-01-01	Protect $500,000 Keep It!	MrBeast	Entertainment	85458562	3598247	4.21
7	JP	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	137639799	4438394	3.22

Question 4: 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? Only look at the data from 2022. Order the result by category_title and country.

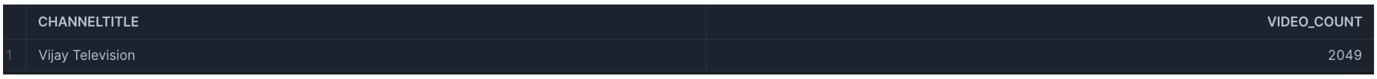
The query identifies the top category by distinct video titles for each country from 2022 onwards. It calculates the percentage of these distinct titles within the total distinct videos for that country, rounded to two decimal places. The result is ordered by category title and country, showing that in most countries, the "Entertainment" category had the most distinct videos, while in Canada and the US, "Gaming" dominated.



	COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
1	BR	Entertainment	5417	23760	22.80
2	DE	Entertainment	7709	30719	25.10
3	FR	Entertainment	7548	32849	22.98
4	GB	Entertainment	5643	27855	20.26
5	IN	Entertainment	21281	50250	42.35
6	JP	Entertainment	5658	17627	32.10
7	KR	Entertainment	5122	15175	33.75
8	MX	Entertainment	4195	17532	23.93
9	CA	Gaming	6594	30869	21.36
10	US	Gaming	6226	28799	21.62

Question 5: Which channel title has produced the most distinct videos and what is this number?

The answer for question 5 was Vijay Television.



	CHANNELTITLE	VIDEO_COUNT
1	Vijay Television	2049

5. Business Question

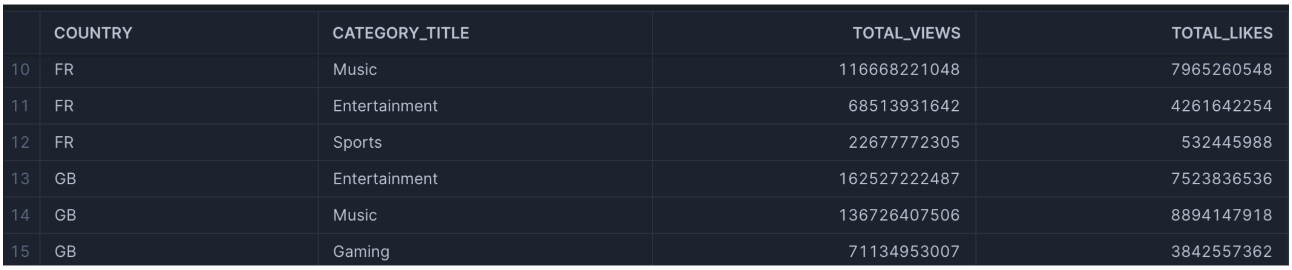
The merged dataset "table_youtube_final" was analysed to find the category with the highest potential to top the charts in as many countries as possible. In our analysis, the entertainment and the music industry has been ignored to ensure ease of content creation and lesser copyrights requirements.

A few different analysis were conducted to find the optimal category to create content for. The main parameter that was analysed is the overall number of views. This is because view count is the only parameter that matters while monetising off the platform. Likes, comments etc. increase engagement to widen your audience, thereby increasing the view count. The answer for this question has been found to be the **Gaming Category.**

Queries:

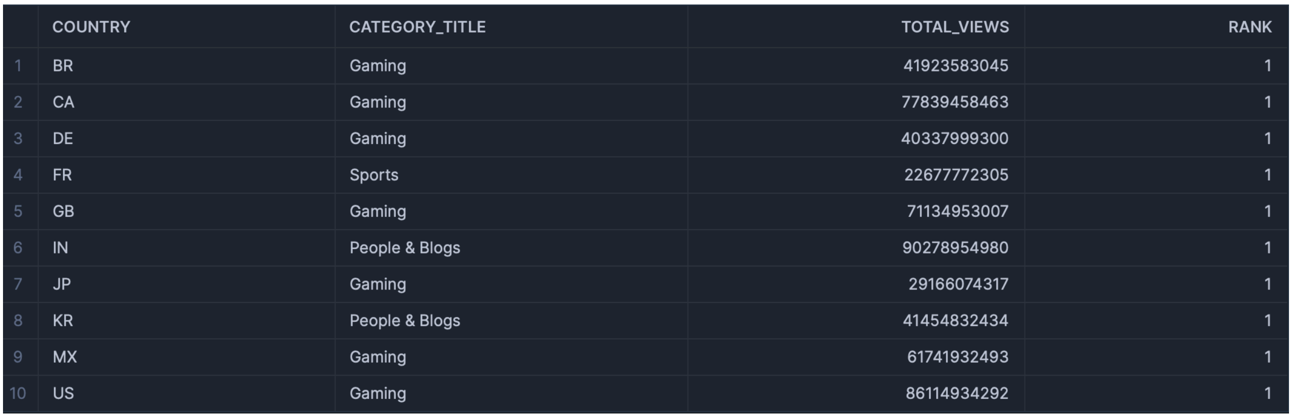
1. At first, the top 3 ranking categories in all the 10 countries have been looked into. A sample of the same has been attached below. It was, as expected, found that music and entertainment rank first and second in every country, with varying orders. We see that the 3rd spot varies in

each country but the gaming category appears in the vast majority. A snippet of the findings has been attached below. **It is key to note that music and entertainment ranked 1st and 2nd in every country.**



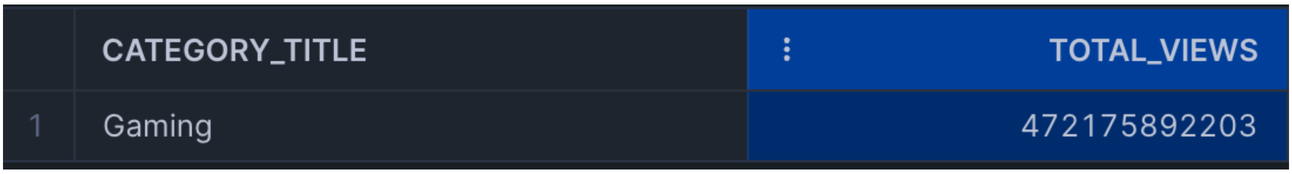
	COUNTRY	CATEGORY_TITLE	TOTAL_VIEWS	TOTAL_LIKES
10	FR	Music	116668221048	7965260548
11	FR	Entertainment	68513931642	4261642254
12	FR	Sports	22677772305	532445988
13	GB	Entertainment	162527222487	7523836536
14	GB	Music	136726407506	8894147918
15	GB	Gaming	71134953007	3842557362

2. Since gaming was seen in most countries, this category was zoomed into and an sql command was generated to find the rank of gaming in each country. The results gave us more clarity on why gaming needs to be the category of interest. While ignoring music and entertainment, it ranked the highest in 7 out of the 10 countries of interest. The findings are attached below.



	COUNTRY	CATEGORY_TITLE	TOTAL_VIEWS	RANK
1	BR	Gaming	41923583045	1
2	CA	Gaming	77839458463	1
3	DE	Gaming	40337999300	1
4	FR	Sports	22677772305	1
5	GB	Gaming	71134953007	1
6	IN	People & Blogs	90278954980	1
7	JP	Gaming	29166074317	1
8	KR	People & Blogs	41454832434	1
9	MX	Gaming	61741932493	1
10	US	Gaming	86114934292	1

3. As one more verification, a sql query was generated to find the category with the highest views while ignoring music and entertainment. The result was again the gaming category with an overall view count as given below.



	CATEGORY_TITLE	TOTAL_VIEWS
1	Gaming	472175892203

Based on all the queries given above, the category that will be finalised on is the gaming category as it shows maximum potential for topping the charts with its views.

6. Challenges & Learnings

The data engineering assignment using Azure and Snowflake came with its challenges and key learning points which have been highlighted below.

6.1 Challenges:

Some of the challenges faced were as follows:

- **Inconsistencies with Large Datasets:** Working with and handling a large dataset came with its challenges such as identifying inconsistencies, missing values, duplicates and so on, which needed careful analysis to ensure minimal noise exists in the dataset before going into the analysis portion of the project.

- **Complexity of SQL queries:** The SQL queries required for some of the analysis conducted required the use of more complex and comprehensive techniques such as the use of Common Table Expressions (CTEs) and functions such as row_number().

- **Correct Interpretation of results:** Working with large datasets also requires extremely careful evaluation of the features and parameters used. Receiving an output from the query doesn't necessarily mean that it is the desired output due to the numerous factors that exist within these large datasets.

- **Proper external and internal table creation:** The creation of internal and external tables posed a challenge with overall understanding of its meaning, significance, and requirement. Ensuring all fields and records were imported, and making sure they were of the required datatype posed a challenge due to the varying datatypes they had been stored in originally.

- **Date and time manipulations:** Aggregating data from various date time formats to extract and filter data from a certain timeframe posed a challenge.

6.2 Key Learnings

- **The usage of new tools and techniques:** The main learning from this project was the understanding of how a data lake works through the use of Microsoft Azure and Snowflake. Working with these tools and techniques shed a light on how large datasets are managed and used within organisations.
- **Advanced SQL techniques:** Understanding was gained on how important and efficient some of the more advanced SQL techniques are and how they can be leveraged to give us key insights on the dataset of interest. The understanding on the use of CTE's row number, truncate for date time, etc. were some of the techniques that were refreshed on and used in a much larger setting than before.

- **Data Analysis best practice:** Working with these large datasets also shed a light on the importance of code structure and clarity. The importance of ensuring proper indentation, spacing, comments, and so on were understood and implemented.

- **Debugging:** Another key learning in this project was the debugging factor. As touched upon before, with large datasets the result that has been achieved through the code is not necessarily the result that we may be after. Using various debugging techniques, learning new and alternative SQL methods to use, and gaining a deeper understanding of the dataset to do so was another key learning.

- **Understanding Business Impact:** Through the use of data driven decision making, first hand experience was gained on how business decisions can be driven through proper use of SQL queries to derive insights from these large oceans of data.