DBS Project 1

Jack Luna

October 2022

1 Introduction

For my data in this project, I decided to look at data regarding videos on youtube. Specifically videos on the trending page, because that's the best dataset I found. I chose this topic because I have some experience working with youtube content, and I've always been curious about the data behind it. My ultimate aim is to analyze what factors contribute to a video on youtube performing well or poorly.

2 The Data

My dataset is a subset of the Trending Youtube Video Statistics dataset from kaggle, found here: https://www.kaggle.com/datasets/datasnaek/youtube-new

The original dataset is unfortunately too large to include in the submission on github, even zipped, so find it at the link.

The dataset covers many countries, but I pared it down to the USA, Canada, and Britain to minimize difficulty with non-english characters and languages I don't know. For some questions, where I thought it was relevant, I considered all three countries. For others, I only considered the USA. The fourth table in this database is an index of video categories. The original dataset has one for each reigon, since the categories are different in some countries, but the three countries in question have the same categories so I only included one category index table in the database.

The category index was initially a .json file, so I used a python script (also included in the distribution if you want it) to convert it to a csv I could load as a table.

3 Queries

Below are the insightful questions I settled on for my data. Along with each question is it's motive, relational algebra representation, and SQLite query, as

well as the results of the queries and brief analyses.

3.1 What are the top categories in each country by video count?

Motive: I'm interested in this question because it can tell us what types of videos are most common in each major market I'm looking at, which can help identify the sentiment of these reigonal audiences.

Relational algebra:

$$_{title}\mathscr{F}_{COUNT(title)}(USvideos \bowtie_{USvides.category\_id=Category.category\_id} Category)$$

SQL:

```
SELECT c.title, Count(c.title) AS "USA Count" FROM USvideos v
    JOIN Category c ON v.category_id = c.id
    GROUP BY c.title ORDER BY Count(c.title) DESC LIMIT 5;
```

Results:

	title	British Count
1	Music	13754
2	Entertainment	9124
3	People & Blogs	2926
4	Film & Animation	2577
5	Howto & Style	1928

	title	Canada Count
1	Entertainment	13451
2	News & Politics	4159
3	People & Blogs	4105
4	Comedy	3773
5	Music	3731

	title	USA Count
1	Entertainment	9964
2	Music	6472
3	Howto & Style	4146
4	Comedy	3457
5	People & Blogs	3210

Insights/notes: The "USvideos" table reference should be replaced with GBvideos or CAvideos in order to check each country's top five categories. The results indicate that, while there are some commonalities, the top five most common video categories DO differ by country. The US seems to trend more entertainment/comedy videos, whereas Canada trends more news and politics higher, and the British trend more Music and animation videos.

3.2 What are the top categories in each country average views per video

Motive: Where the previous section tells us the most commonly posted video types, this question can tell us which types of video actually perform well in a given reigon.

Relational algebra:

$$_{title}\mathscr{F}_{AVG(USvideos.views)}(USvideos \bowtie_{USvides.category\_id=Category.category\_id} Category)$$

SQL:

```
SELECT c.title, AVG(v.views) AS "Average views" FROM USvideos v
    JOIN Category c ON v.category_id = c.id
    GROUP BY c.title ORDER BY AVG(v.views) DESC LIMIT 5;
```

Results:

	title	Average views (BRITISH)
1	Music	12444442.6907809
2	Nonprofits & Activism	3919980.68888889
3	Entertainment	3264607.96153003
4	Film & Animation	3245132.95537447
5	Science & Technology	3168703.1969112

	title	Average views (CANADA)
1	Music	3532524.84427767
2	Movies	2853415.0
3	Nonprofits & Activism	1562184.09459459
4	Film & Animation	1426728.56504854
5	Science & Technology	1233844.52034632

	title	Average views (USA)
1	Music	6201003.11959209
2	Film & Animation	3106250.20085288
3	Nonprofits & Activism	2963884.07017544
4	Gaming	2620830.63035496
5	Entertainment	2067883.19901646

Insights/notes: This metric makes all three countries seem more similar. The one real exception is that the USA's trending page gives more views to Gaming videos, whereas the other two give more to Science and Technology videos. It seems that despite the disparities in what content is produced, all three audiences consume broadly the same type of content.

3.3 What is the best time of day to post a video?

Motive: It's well known in online circles that posting a video at certain times of day can effect it's performance. To analyze this, we can average the views and likes of videos in each hour-long block to see which hour happens to lead to the best performance.

Relational algebra:

$$_{hour\_of\_day*}\mathscr{F}_{AVG(views),AVG(likes)}(USvideos)$$

* where hour_of_day is the 24hr time.

SQL:

```
SELECT strftime('%H', publish_time) AS OffsetHours,
    CAST(AVG(views) AS INTEGER) AS "US Views", CAST(AVG(likes) AS INTEGER) AS "US Likes" FRO
    GROUP BY strftime('%H', publish_time);
```

Results (next page):

	OffsetHours	GB Views	GB Likes
1	00	9039659	190684
2	01	5259994	95553
3	02	4177266	68300
4	03	4370840	90240
5	04	14318343	303842
6	05	14815357	245278
7	06	2368193	43061
8	07	3734901	64531
9	08	6673698	147843
10	09	5897098	277002
11	10	7444624	118779
12	11	6966888	122590
13	12	5733506	138050
14	13	4540890	107350
15	14	5928744	124305
16	15	4774160	159746
17	16	4831191	126424
18	17	4243366	97508
19	18	2858680	97365
20	19	2209379	59144
21	20	4166527	74509
22	21	4576021	129935
23	22	6926614	114627
24	23	6421100	116915

	OffsetHours	CA Views	CA Likes
1	00	982329	36876
2	01	1073116	30390
3	02	893601	23261
4	03	1146501	30408
5	04	2158561	76158
6	05	2563636	88170
7	06	1488243	30670
8	07	1672217	33730
9	08	1689870	44679
10	09	2518229	124519
11	10	975775	24262
12	11	1163467	30071
13	12	1471429	45743
14	13	1608767	48023
15	14	1259075	39940
16	15	1091099	43744
17	16	765200	29804
18	17	1049136	39665
19	18	1027851	41839
20	19	832258	34233
21	20	923697	31503
22	21	966298	35866
23	22	949538	37367
24	23	752726	27415

	OffsetHours	US Views	US Likes
1	00	1464088	50884
2	01	1978763	56413
3	02	2097509	48688
4	03	2884458	72177
5	04	7343508	217217
6	05	2384607	84128
7	06	1642173	33710
8	07	4147581	93571
9	08	3409047	88569
10	09	5748058	264475
11	10	4618096	85245
12	11	2584947	77689
13	12	3114149	78254
14	13	2822246	74302
15	14	2513804	72467
16	15	1990367	85409
17	16	1970129	67420
18	17	1603781	52217
19	18	1756689	54136
20	19	1467202	57757
21	20	1985574	58025
22	21	2090362	78681
23	22	1800147	64945
24	23	1657133	48286

Insights/notes: The times for this are all in UTC-0 24 hour time, because that's what the dataset records all the times in. The best times vary a little by country, but all three have a sweet spot around 0400 UTC, and another around 0900 UTC. The fact that three countries have similar sweet spots, even though one country is in a completely different time zone, suggests that the algorithm that determines which videos make it to trending probably takes into account more than just people from that one country, otherwise the british hotspots would be drastically different to the US and Canada hotspots throughout the day.

3.4 What is the best month to post a video?

Motive: The results of the previous question made me wonder whether there were "best posting times" on a yearly scale, not just daily. So I repeated a similar query, but taking months instead of hours.

Relational algebra:

$$_{month\_of\_year*}\mathscr{F}_{AVG(views),AVG(likes)}(USvideos)$$

* where month_of_year is the 24hr time.
SQL:

```
SELECT strftime('%m', publish_time) AS Month,
    CAST(AVG(views) AS INTEGER), CAST(AVG(likes) AS INTEGER) FROM USvideos
    GROUP BY strftime('%m', publish_time);
```

Results:

	Month	CAST(AVG(views) AS INTEGER)	CAST(AVG(likes) AS INTEGER)
1	01	1112944	42980
2	02	1447107	41439
3	03	2089192	71046
4	04	3894051	108283
5	05	4871875	142175
6	06	3356845	139873
7	07	41019	87
8	08	79462	1189
9	09	18905	92
10	10	69662	1061
11	11	1323492	49000
12	12	1314857	47489

Insights/notes: I stuck to just the US times for this one, since it didn't seem like reigon would matter as much over a year period. One difficulty with this query is that this data only covers trending page info from November 2017 to July 2018, so several months, (the ones with the noticably lower numbers in the results), are not properly accounted for. Of the months we actually have proper data for, it seems like the summer months are the highest-performing times to post. This makes a lot of sense, considering kids are on summer break and probably watch a lot more content during that time.

3.5 Which channels were most trending in the time period represented by the data? Does this differ from country to country?

Motive: This can tell us which creators trend most often per country, which can give us further info on the leaning of the audiences in each country.

Relational algebra:

$$_{channel\_title}\mathscr{F}_{COUNT(channel\_title)}(USvideos)$$

SQL:

```
SELECT channel_title, COUNT(channel_title) FROM USvideos
    GROUP BY channel_title ORDER BY COUNT(channel_title) DESC LIMIT 10;
```

Results:

	channel_title	GB count
1	The Tonight Show Starring Jimmy Fallon	208
2	TheEllenShow	207
3	Jimmy Kimmel Live	207
4	Saturday Night Live	206
5	WWE	205
6	The Late Late Show with James Corden	202
7	Late Night with Seth Meyers	194
8	Breakfast Club Power 105.1 FM	193
9	The Late Show with Stephen Colbert	189
10	Netflix	187

	channel_title	CA count
1	SET India	192
2	MSNBC	189
3	FBE	188
4	The Young Turks	186
5	REACT	183
6	VikatanTV	182
7	CNN	182
8	The Late Show with Stephen Colbert	172
9	RadaanMedia	168
10	ARY Digital	168

	channel_title	US count
1	ESPN	203
2	The Tonight Show Starring Jimmy Fallon	197
3	Vox	193
4	TheEllenShow	193
5	Netflix	193
6	The Late Show with Stephen Colbert	187
7	Jimmy Kimmel Live	186
8	Late Night with Seth Meyers	183
9	Screen Junkies	182
10	NBA	181

Insights/notes: To a pretty big extent, these results reflect the results of the top categories questions above. The US favors entertainment (Netflix, ESPN,

Screen Junkies), while Britain HEAVILY favors late night talk show channels, and Canada favors news (MSNBC, SET India, CNN) and politics (The Young Turks). There are fewer similarities here than in the category questions, which suggests that the three countries are more distinct in who they sit down to watch, even if the broad categories have more of an overlap.

3.6 What proportion of trending videos have comments disabled? How did those videos perform relative to comments-enabled videos?

Motive: This was mostly a curiosity from me. The dataset offered info on whether the videos had comments enabled, and I wanted to find out if comment-disabling correlated to a change in performance. My thought process was that generally only controversial videos disable their comments, so there might be a correlation there.

Relational algebra:

$$_{comments\_disabled}\mathscr{F}_{COUNT(comments\_disabled),AVG(views),AVG(likes)}(USvideos)$$

SQL:

```
SELECT comments_disabled, COUNT(comments_disabled), CAST(AVG(views) AS INTEGER) AS Views,
    CAST(AVG(likes) AS INTEGER) AS Likes FROM USvideos GROUP BY comments_disabled;
```

Results:

	comments_disabled	COUNT(comments_disabled)	Views	Likes
1	False	40316	2358304	75096
2	True	633	2518743	21445

Insights/notes: The results here are actually fairly interesting! The average number of views on comments-disabled videos is roughly equal to those of non-disabled videos, but the number of likes tend to be much lower! This suggests that my controversial video hypothesis may be accurate. Plenty of people still watch videos without comment sections, but WAY fewer of them are found enjoyable.

3.7 What relationship, if any, exists between title length and video performance?

Motive: Titles of youtube videos are often talked about as a key component in getting people to actually watch what you post. My hypothesis going into

this one was that shorter, more eye-friendly titles would draw more viewership than extremely long titles.

Relational algebra:

$$_{title}\mathscr{F}_{length(title),AVG(views),AVG(likes)}(USvideos)$$

SQL:

```
SELECT * FROM (
    SELECT title, length(title) AS lengthTitle, CAST(AVG(views) AS INTEGER) AS Views,
    CAST(AVG(likes) AS INTEGER) AS Likes
    FROM USvideos GROUP BY title ORDER BY length(title) DESC LIMIT 10
)
UNION
SELECT * FROM (
    SELECT title, length(title) AS lengthTitle, CAST(AVG(views) AS INTEGER) AS Views,
    CAST(AVG(likes) AS INTEGER) AS Likes
    FROM USvideos GROUP BY title ORDER BY length(title) ASC LIMIT 10
)
ORDER BY lengthTitle;
```

Results:

	title	lengthTitle	Views	Likes
1	435	3	1840254	121698
2	Jack	4	3065820	253776
3	love	4	277071	20994
4	Voices	6	111673	1124
5	*cough*	7	2122477	216503
6	Iceberg	7	124022	212
7	Kittens	7	919012	58047
8	Bye 2017	8	257111	25213
9	Patience	8	1987167	132185
10	Spinners	8	665457	18913
11	Patrick Beverley calls in to tell Will he knows ...	100	261660	3300
12	Revlon Live Boldly | Anthem with Ashley Graham,...	100	7130	114
13	Ronaldo knocks out Juventus with last minute ...	100	1193011	8233
14	Rony Abovitz, founder of Magic Leap, and Adam ...	100	33854	181
15	Stephen A. sides with Reggie Miller: Magic ...	100	650368	4666
16	Stephen A.: Cavs 'were an absolute disgrace' in ...	100	476811	4155
17	Stephen A.: Kevin Durant looked like he wanted ...	100	978831	8211
18	The Twelfth Doctor Regenerates: Peter Capaldi t...	100	1200785	16324
19	The epic late-night Fortnite stream featuring ...	100	410884	7122
20	U2: Live in der Berliner U-Bahnlinie U2: „Get Out ...	100	47117	444

Insights/notes: For the actual SQL here, I cut out the middle and just looked at the shortest and longest titles. They seem to somewhat confirm my hypothesis, but the data is unclear. If I could do this again I'd get more nuanced data than just the very shortest and very longest titles. As it is, it seems like the shortest video titles correspond to some of the highest AND some of the lowest view counts, whereas the long titles tend to have more moderate performance.

3.8 Do heavily liked videos or heavily disliked videos get more comments?

Motive: While thinking about controversial videos for a previous question, I decided to check on that issue more explicitly. This query computes a ratio of likes to dislikes

Relational algebra:

$$_{title}\mathscr{F}_{likes/dislikes,AVG(views)}(USvideos)$$

SQL:

```
SELECT * FROM (
    SELECT title, views, likes, dislikes, likes/dislikes AS ratio, comment_count
    FROM USvideos WHERE likes > 10000 AND dislikes > 10000
    GROUP BY title ORDER BY likes/dislikes DESC
    LIMIT 10
)
UNION
SELECT * FROM (
    SELECT title, views, likes, dislikes, likes/dislikes AS ratio, comment_count
    FROM USvideos WHERE likes > 10000 AND dislikes > 10000
    GROUP BY title ORDER BY likes/dislikes AS    LIMIT 10
)
ORDER BY ratio;
```

Results:

	title	views	likes	dislikes	ratio	comment_count
1	#MeToo Backlash | January 17, 2018 Act 1 | Full...	635686	11383	12122	0	4289
2	#ProudToCreate: Pride 2018	597669	29781	71617	0	55110
3	DAD TEACHES GAY SON HOW TO SHOOT ...	408504	10268	13209	0	5136
4	Emma Gonzalez gives speech at March for Our ...	486688	11540	12891	0	16387
5	Fergie Performs The U.S. National Anthem / 201...	14647590	32892	117128	0	44404
6	Fergie's national anthem draws criticism | ESPN	2762839	14362	37877	0	21507
7	I Feel Pretty Trailer #1 | Movieclips Trailers	923181	10099	36196	0	7115
8	Matthew Santoro - FACTS (Official Music Video) ...	328330	15186	15448	0	7484
9	Our Surrogate Story: The Truth | Gigi	759844	18096	21877	0	4477
10	Why black Americans are getting less sleep	664464	10001	34147	0	10818
11	Ed Sheeran - Perfect (Official Music Video)	33523622	1634124	21082	77	85067
12	MEDICINE - QUEEN NAIJA (OFFICIAL VIDEO)	15797464	788034	10012	78	128541
13	Marvel Studios' Avengers: Infinity War Official ...	37736281	1735895	21969	79	241237
14	BTS (방탄소년단) 'MIC Drop (Steve Aoki Remix)' ...	13945717	2055137	23888	86	395562
15	GOT7 Look M/V	21481226	987147	10106	97	195966
16	j-hope 'Daydream (백일몽)' MV	10695328	2050527	14711	139	387384
17	BTS (방탄소년단) LOVE YOURSELF 轉 Tear ...	10666323	1956202	13966	140	285583
18	BTS (방탄소년단) 'Euphoria : Theme of LOVE ...	10208424	1795542	10956	163	201589
19	BTS (방탄소년단) 'FAKE LOVE' Official MV ...	10498453	1758871	10522	167	152997
20	j-hope 'Airplane' MV	9879448	1731826	10227	169	176698

Insights/notes: These are more concrete results. While the disliked videos do get more viewership than you would expect from it's near-zero (so close it's rounded off) like-dislike ratio, highly-liked videos almost universally get two degrees of magnitude more views by this data. All press may be good press, but positive press is still apparently better than negative press!

3.9 How does number/length of tags effect performance?

Motive: The tags in a video's description are another talked-about topic in the realm of getting people to watch your videos. I have no way of analyzing what the tags are with SQL alone, but one thing I can check is whether the number of tags you choose effects your video's virality!

Relational algebra:

$$_{num\_tags}\mathscr{F}_{num\_tags,likes}(USvideos)$$

I'm not sure about this relational algebra, see SQL below for implementation. Num_tags is a calculated quantity based on the occurences of the tag delimiter in the tags column.

SQL:

```
SELECT * FROM (
    SELECT (LENGTH(tags) - LENGTH(REPLACE(tags,'|', '')))+1 AS num_tags, views, likes FROM U
    BY (LENGTH(tags) - LENGTH(REPLACE(tags,'|', '')))+1 DESC LIMIT 10
)
UNION
SELECT * FROM (
    SELECT (LENGTH(tags) - LENGTH(REPLACE(tags,'|', '')))+1 AS num_tags, views, likes FROM U
    BY (LENGTH(tags) - LENGTH(REPLACE(tags,'|', '')))+1 ASC LIMIT 10
)
ORDER BY num_tags;
```

Results:

	num_tags	**views**	likes
1	1	945	7
2	1	6412	49
3	1	27943	156
4	1	95944	1354
5	1	219030	14303
6	1	284666	16396
7	1	366180	4364
8	1	748374	57527
9	1	762616	20159
10	1	1624771	84426
11	69	21798	1268
12	69	23437	1301
13	69	24413	1334
14	69	25130	1345
15	69	432987	6751
16	69	765238	8818
17	69	910538	9883
18	69	961176	10363
19	69	981210	10523
20	69	997733	10642

Insights/notes: It seems that low numbers of tags perform about as well as high numbers. There isn't much of a difference there.