Data collection from the Youtube API, cloud databases and EDA

Agenda

1. Introduction

2. Objectives

3. Youtube API

4. Database

5. Results and Discussion

6. Conclusions

1. Introduction

In this project, Grubby's youtube channel was analyzed in terms of views, likes, dislikes and comments on its videos about the new game Age of Empires IV. Data was collected by the Youtube API and was placed in a PostgreSQL database in the cloud on a server of Amazon Web Services (AWS). The whole project was done in the Python programming language. The methods and code of how the data was obtained and allocated in the database will be exposed later. Finally, an exploratory data analysis (EDA) was performed using Power BI.

Age of Empires IV is a real-time strategy (RTS) video game developed by Relic Entertainment in partnership with World's Edge and published by Xbox Game Studios. The game was released on October 28th, 2021. The game drew a lot of attention as it is the only triple RTS game developed and released since Starcraft 2, which was released in 2010 by Blizzard Entertainment. Its success is demonstrated in its numbers, all-time peak players last month 73,928 (steamdb), 19,162 very positives reviews on steam and thousands of games played every day.

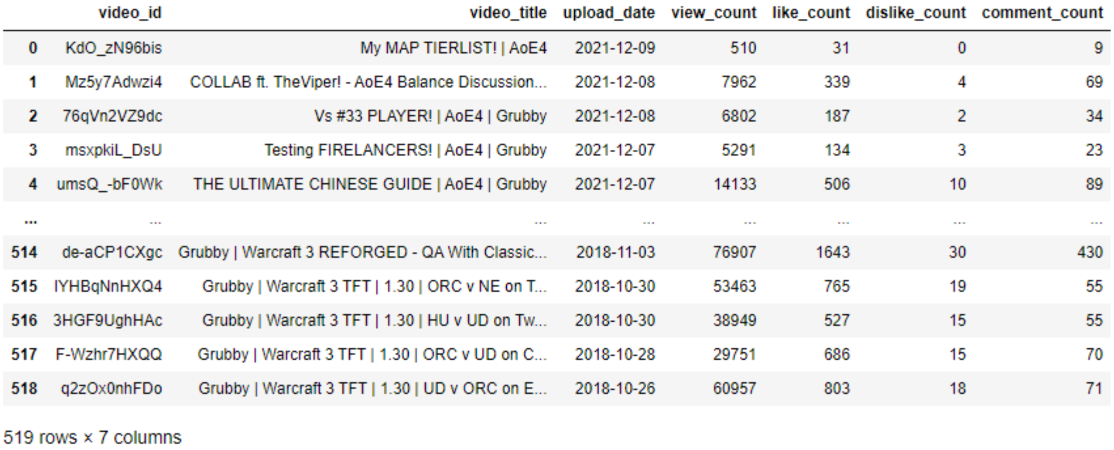
Manuel Schenkhuizen, also known as Grubby, is a Dutch real-time strategy gamer and former professional esports player. He competed in the RTS games Warcraft III (WC3), Warcraft III: The Frozen Throne and Starcraft II. Grubby is the most successful WC3 player of all time, as an Orc player, having won more than 38 LAN tournaments (in person events), of which six were World Championships (WIKIPEDIA). Now he livestreams at the livestream platform Twitch and he has a youtube channel (FollowGrubby) with more than 1 million views, 227,000 subscribers and 500 videos published to this date.

All the code can be found in the link of my github at the end of the post.

1. Objectives

The objective of this project was to put into practice the use of one of the best technologies for obtaining data currently employed by most companies, the APIs, and the use of the cloud database to store the data obtained, perform EDA in the collected data, and analyze which videos performed best and their motives.

2. Youtube API



	video_id	video_title	upload_date	view_count	like_count	dislike_count	comment_count
0	KdO_zN96bis	My MAP TIERLIST! | AoE4	2021-12-09	510	31	0	9
1	Mz5y7Adwzi4	COLLAB ft. TheViper! - AoE4 Balance Discussion...	2021-12-08	7962	339	4	69
2	76qVn2VZ9dc	Vs #33 PLAYER! | AoE4 | Grubby	2021-12-08	6802	187	2	34
3	msxpkiL_DsU	Testing FIRELANCERS! | AoE4 | Grubby	2021-12-07	5291	134	3	23
4	umsQ_-bF0Wk	THE ULTIMATE CHINESE GUIDE | AoE4 | Grubby	2021-12-07	14133	506	10	89
...	...	...	...	...	...	...	...
514	de-aCP1CXgc	Grubby | Warcraft 3 REFORGED - QA With Classic...	2018-11-03	76907	1643	30	430
515	lYHBqNnHXQ4	Grubby | Warcraft 3 TFT | 1.30 | ORC v NE on T...	2018-10-30	53463	765	19	55
516	3HGF9UghHAc	Grubby | Warcraft 3 TFT | 1.30 | HU v UD on Tw...	2018-10-30	38949	527	15	55
517	F-Wzhr7HXQQ	Grubby | Warcraft 3 TFT | 1.30 | ORC v UD on C...	2018-10-28	29751	686	15	70
518	q2zOx0nhFDo	Grubby | Warcraft 3 TFT | 1.30 | UD v ORC on E...	2018-10-26	60957	803	18	71

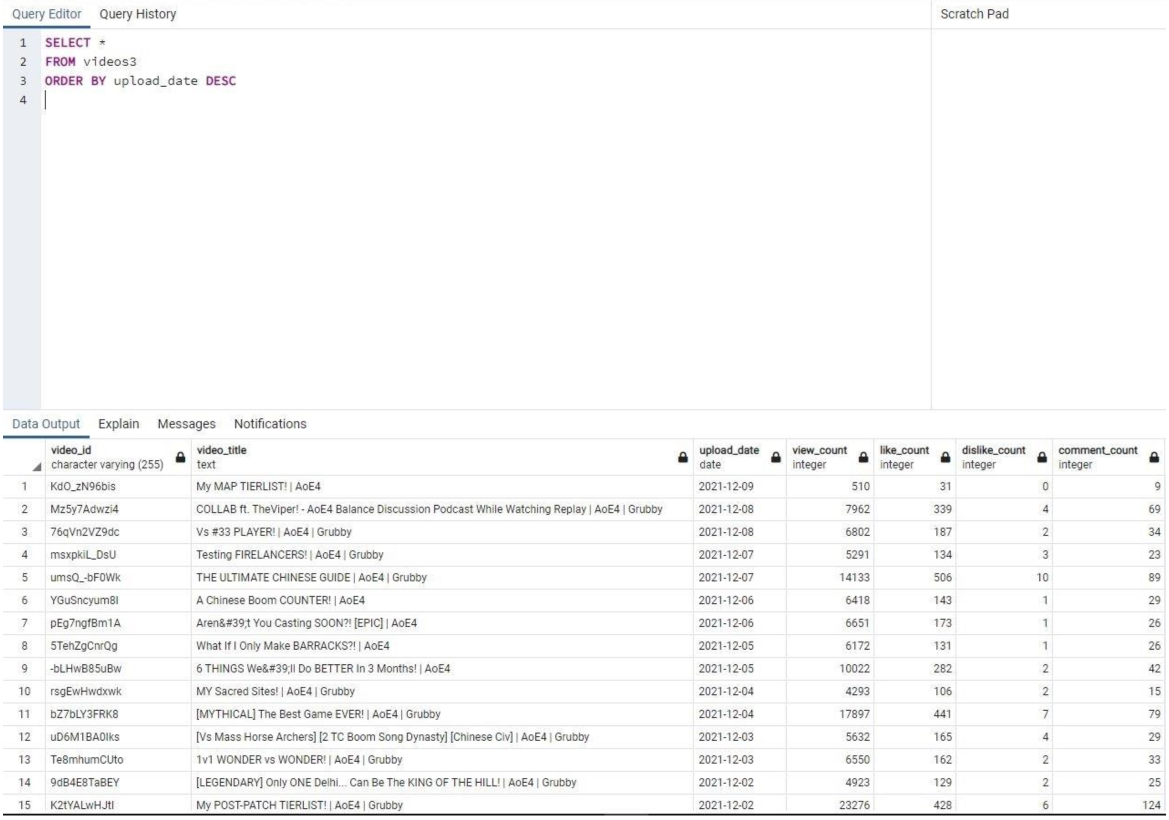
519 rows × 7 columns

Figure 1 - Pandas dataframe with the data collected from the API

The first part of the project was to get the api data from youtube. For this, an API key was created and the channel ID was obtained, essences for the operation of the API. In their possession, an API call is made using the request library and the data is obtained in JSON format. From the API, vídeo_id, vídeo_title, upload_date, view_count, like_count, dislike_count and comment_count were requested. The data has already been pre-requested in order of publication of the videos, from the most recent to the oldest. Then the data was saved in a Panda's dataframe in order to save it in a database, as seen in Figure 1. Finally, the data was saved in a csv file.

3. Database

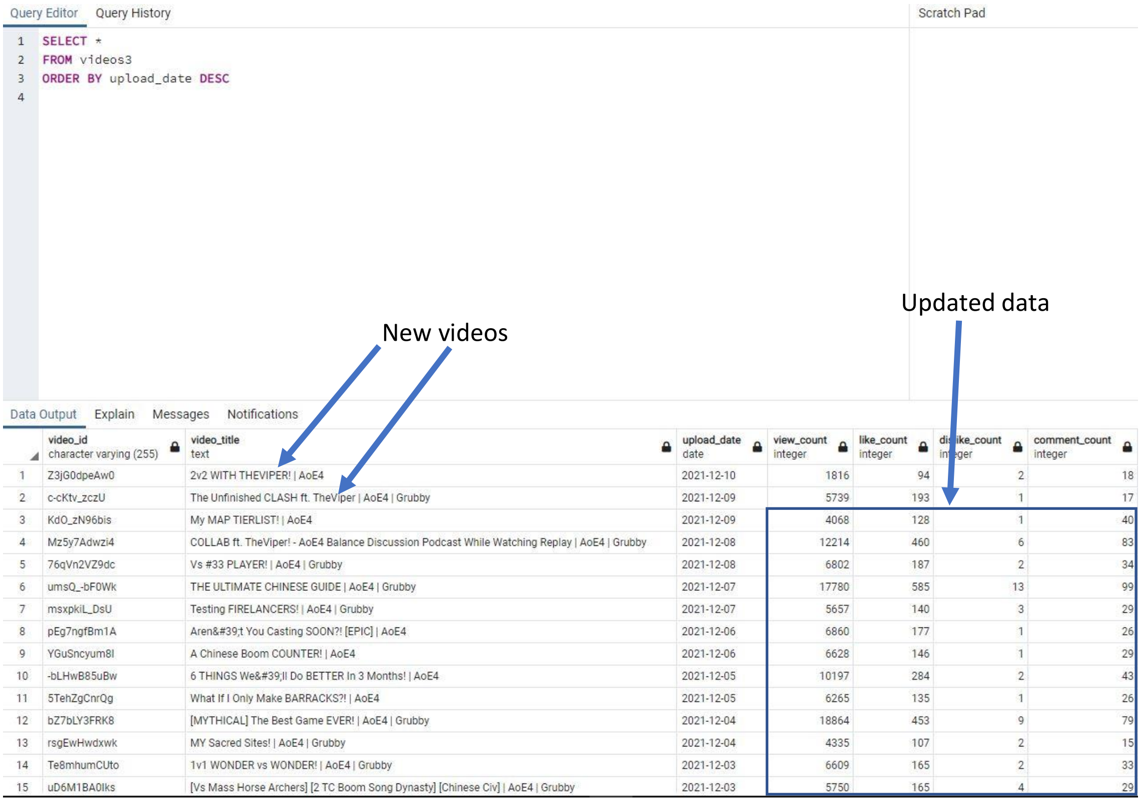
The cloud database chosen was relational type, PostgreSQL, and it was created on the AWS platform. After the connection was created and made to the jupyter notebook, the first commit was made to load the data into the database.



```sql
SELECT *
FROM videos3
ORDER BY upload_date DESC
```

	video_id character varying (255)	video_title text	upload_date date	view_count integer	like_count integer	dislike_count integer	comment_count integer
1	KdO_zN96bis	My MAP TIERLIST! | AoE4	2021-12-09	510	31	0	9
2	Mz5y7Adwzi4	COLLAB ft. TheViper! - AoE4 Balance Discussion Podcast While Watching Replay | AoE4 | Grubby	2021-12-08	7962	339	4	69
3	76qVn2VZ9dc	Vs #33 PLAYER! | AoE4 | Grubby	2021-12-08	6802	187	2	34
4	msxpkiL_DsU	Testing FIRELANCERS! | AoE4 | Grubby	2021-12-07	5291	134	3	23
5	umsQ_-bF0Wk	THE ULTIMATE CHINESE GUIDE | AoE4 | Grubby	2021-12-07	14133	506	10	89
6	YGuSncyum8I	A Chinese Boom COUNTER! | AoE4	2021-12-06	6418	143	1	29
7	pEg7ngfBm1A	Aren&#39;t You Casting SOON?! [EPIC] | AoE4	2021-12-06	6651	173	1	26
8	5TehZgCnrQg	What If I Only Make BARRACKS?! | AoE4	2021-12-05	6172	131	1	26
9	-bLHwB85uBw	6 THINGS We&#39;ll Do BETTER In 3 Months! | AoE4	2021-12-05	10022	282	2	42
10	rsgEwHwdxwk	MY Sacred Sites! | AoE4 | Grubby	2021-12-04	4293	106	2	15
11	bZ7bLY3FRK8	[MYTHICAL] The Best Game EVER! | AoE4 | Grubby	2021-12-04	17897	441	7	79
12	uD6M1BA0Iks	[Vs Mass Horse Archers] [2 TC Boom Song Dynasty] [Chinese Civ] | AoE4 | Grubby	2021-12-03	5632	165	4	29
13	Te8mhumCUto	1v1 WONDER vs WONDER! | AoE4 | Grubby	2021-12-03	6550	162	2	33
14	9dB4E8TaBEY	[LEGENDARY] Only ONE Delhi... Can Be The KING OF THE HILL! | AoE4 | Grubby	2021-12-02	4923	129	2	25
15	K2tYALwHJtI	My POST-PATCH TIERLIST! | AoE4 | Grubby	2021-12-02	23276	428	6	124

Figure 2 - Data successfully loaded to the database



```sql
SELECT *
FROM videos3
ORDER BY upload_date DESC
```

	video_id character varying (255)	video_title text	upload_date date	view_count integer	like_count integer	dislike_count integer	comment_count integer
1	Z3jG0dpeAw0	2v2 WITH THEVIPER! | AoE4	2021-12-10	1816	94	2	18
2	c-cKtv_zczU	The Unfinished CLASH ft. TheViper | AoE4 | Grubby	2021-12-09	5739	193	1	17
3	KdO_zN96bis	My MAP TIERLIST! | AoE4	2021-12-09	4068	128	1	40
4	Mz5y7Adwzi4	COLLAB ft. TheViper! - AoE4 Balance Discussion Podcast While Watching Replay | AoE4 | Grubby	2021-12-08	12214	460	6	83
5	76qVn2VZ9dc	Vs #33 PLAYER! | AoE4 | Grubby	2021-12-08	6802	187	2	34
6	umsQ_-bF0Wk	THE ULTIMATE CHINESE GUIDE | AoE4 | Grubby	2021-12-07	17780	585	13	99
7	msxpkiL_DsU	Testing FIRELANCERS! | AoE4 | Grubby	2021-12-07	5657	140	3	29
8	pEg7ngfBm1A	Aren&#39;t You Casting SOON?! [EPIC] | AoE4	2021-12-06	6860	177	1	26
9	YGuSncyum8I	A Chinese Boom COUNTER! | AoE4	2021-12-06	6628	146	1	29
10	-bLHwB85uBw	6 THINGS We&#39;ll Do BETTER In 3 Months! | AoE4	2021-12-05	10197	284	2	43
11	5TehZgCnrQg	What If I Only Make BARRACKS?! | AoE4	2021-12-05	6265	135	1	26
12	bZ7bLY3FRK8	[MYTHICAL] The Best Game EVER! | AoE4 | Grubby	2021-12-04	18864	453	9	79
13	rsgEwHwdxwk	MY Sacred Sites! | AoE4 | Grubby	2021-12-04	4335	107	2	15
14	Te8mhumCUto	1v1 WONDER vs WONDER! | AoE4 | Grubby	2021-12-03	6609	165	2	33
15	uD6M1BA0Iks	[Vs Mass Horse Archers] [2 TC Boom Song Dynasty] [Chinese Civ] | AoE4 | Grubby	2021-12-03	5750	165	4	29

Figure 3 - Updated data

In Figure 2, we see that all data collected from API, vídeo_id, vídeo_title, upload_date, view_count, like_count, dislike_count and comment_count, were successfully imported. They were organized by date, as seen in the query above the table. To test whether the code is working for when new videos are added to the channel and when the numbers are updated by the API, a second API call and updated new data was made in the database, as seen in Figure 3. In it, we see two new videos, '2v2 WITH THEVIPER' and 'The unfinished CLASH ft TheViper', and also the updated data for all other videos.

1. Results and Discussion

With the database working properly, the next step was to do an analysis of the data. In my other projects I almost always use Python and the matplotlib and seaborn libraries to do the analysis, so I decided to use Power BI this time. All the steps in detail that were done in Power BI, in the 'Transform Data' part, and the previews are available on github as well. In addition to the metrics obtained by the API, I implemented two more metrics for a deeper analysis. They are: 'views per like', how many views are required to generate 1 like and 'views per comment', how many views are required to generate 1 comment. In both we want the smallest possible number, because the smaller these numbers, the fewer views are needed to generate 1 comment or 1 like, generating greater engagement in the videos.

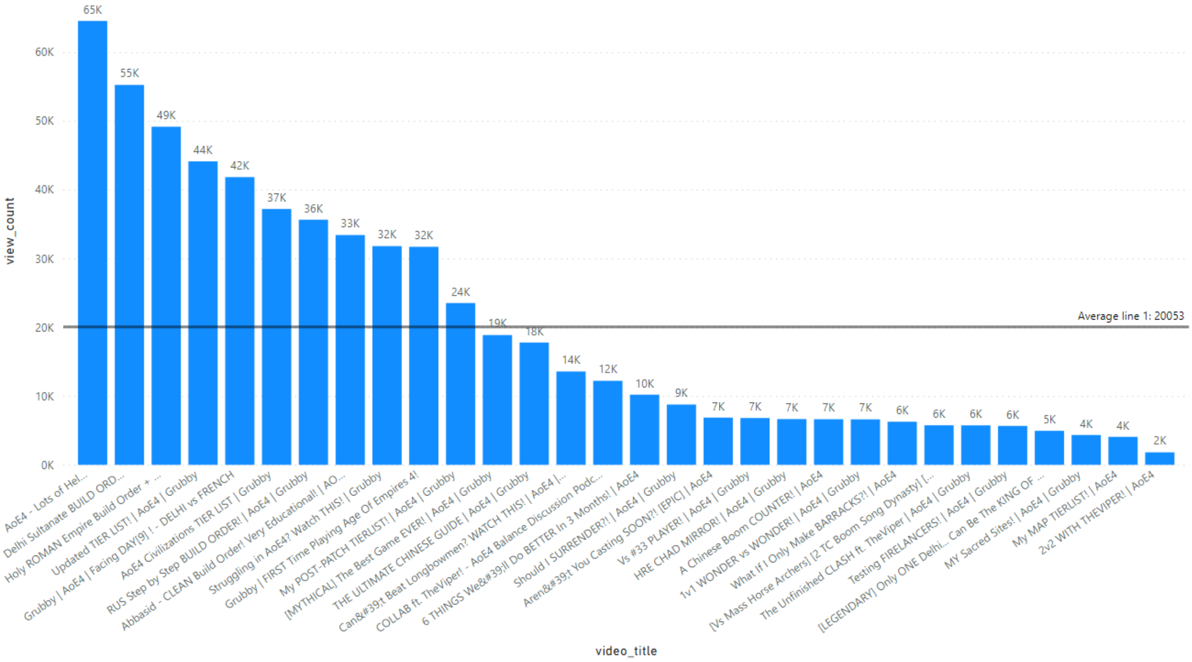


Figure 4 - View_count for each vídeo

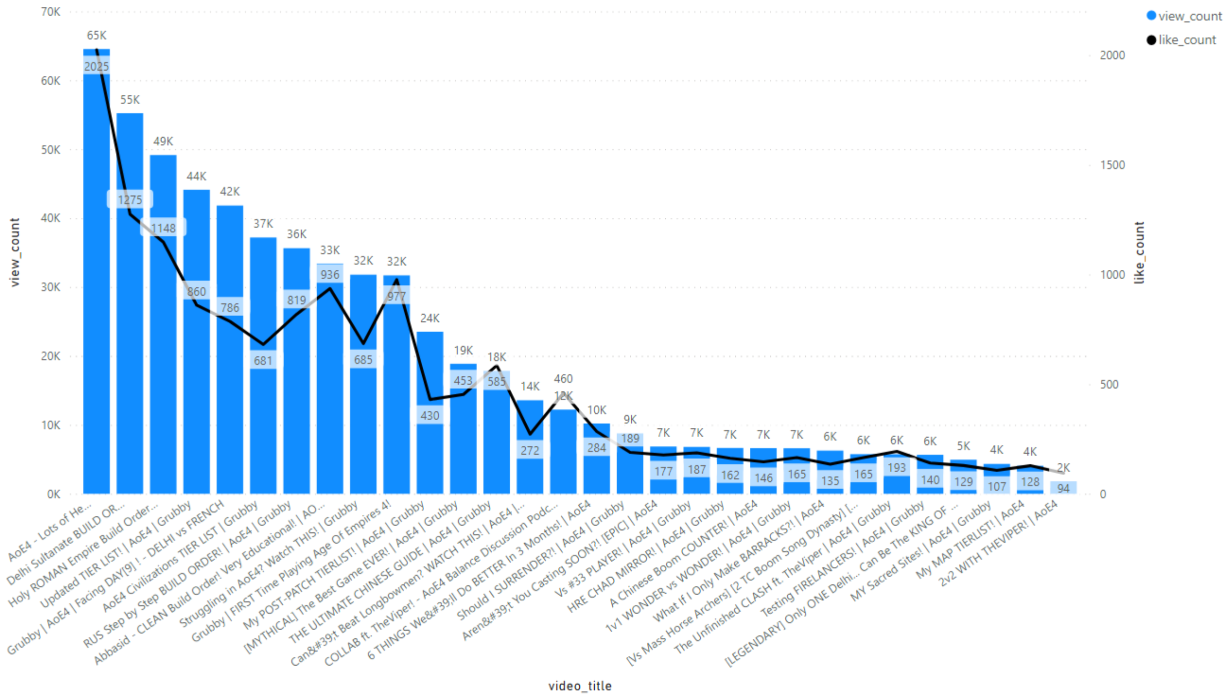


Figure 5 - View_count and like_count for each video

As the scope of the project is the game Age of Empires, the analysis will contain only the videos on this subject. In Figure 4, we see which videos have the highest number of views and the average number of views. Videos containing educational content, which would be videos helping other players improve in the game, have the highest number of views and all are above the midline of views. The other videos, which are mostly gameplay, don't perform as well. This is due to the game being new and many people look for videos on youtube to help them improve. In Figure 5, we have the same views for each video, but with the amount of likes they received. We soon realize that, generally, the more views, the greater the tendency to have more likes, but for example, we have the videos 'AoE4 Civilizations TIER LIST' and 'Struggling in AoE4? Watch THIS!' have not had a similar amount of likes from other videos with the same number of likes.

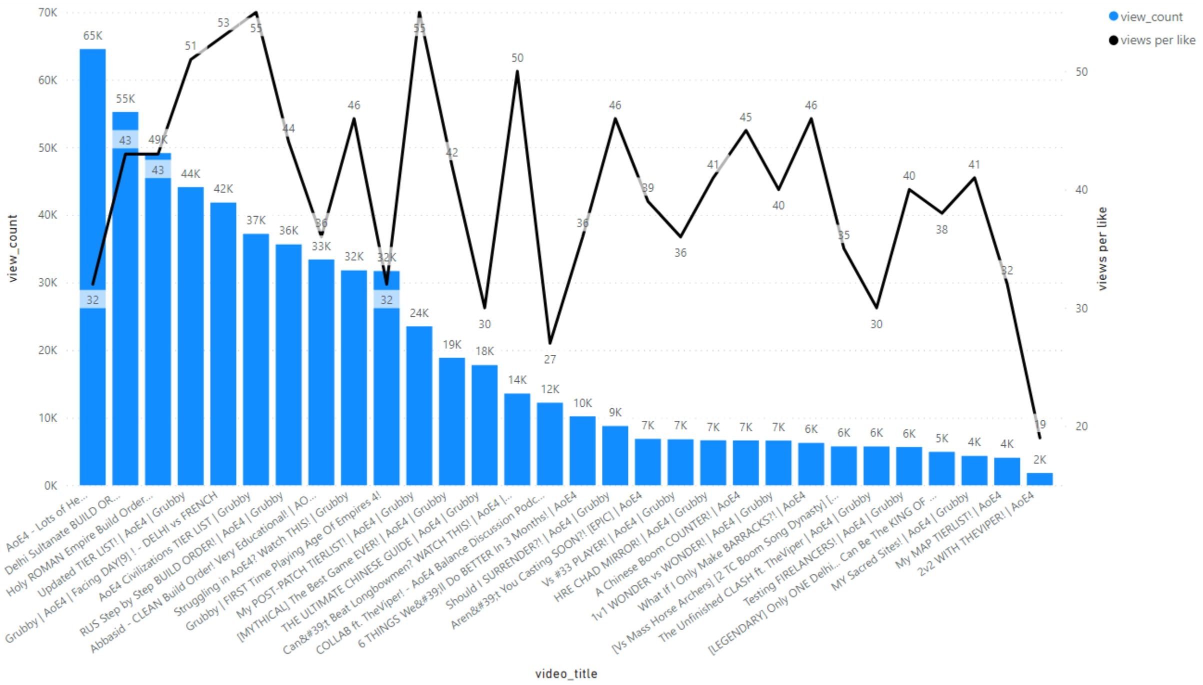


Figure 6 - Views per like

To generate hitching on youtube likes are key and the more likes a video has the better. In Figure 6, we have the required number of views to generate a like. Comparing videos with almost the same number of views, we have those videos with the participation of someone else (collabs) generate more engagement, for example, 'The Unfinished FT CLASH. TheViper' generated more engagement than the video 'What if I only make BARRACKS?', even both. The contents containing how to play guides are also the ones that generate the most engagement, compared to other videos with the number of similar views, for example, 'THE ULTIMATE CHINESE GUIDE' and 'Can't beat longbowmen? WATCH THIS!'.

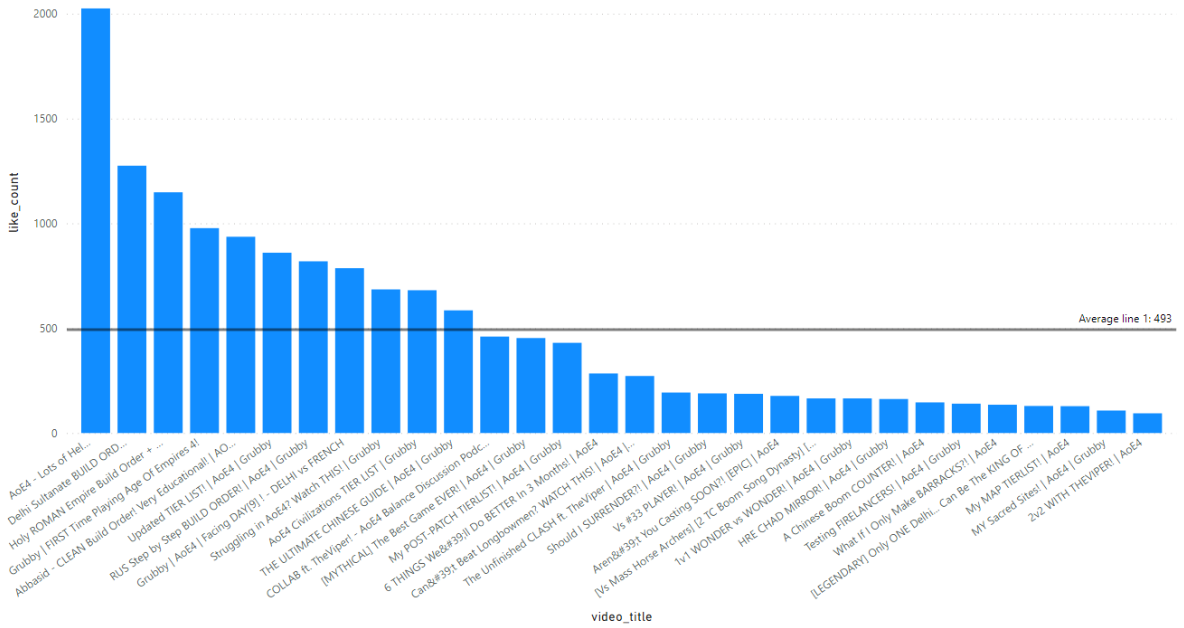


Figure 7 - like_count for each vídeo

Figure 7 confirms that the category of guides is always well received by viewers and generates an engagement, in the likes, above average.

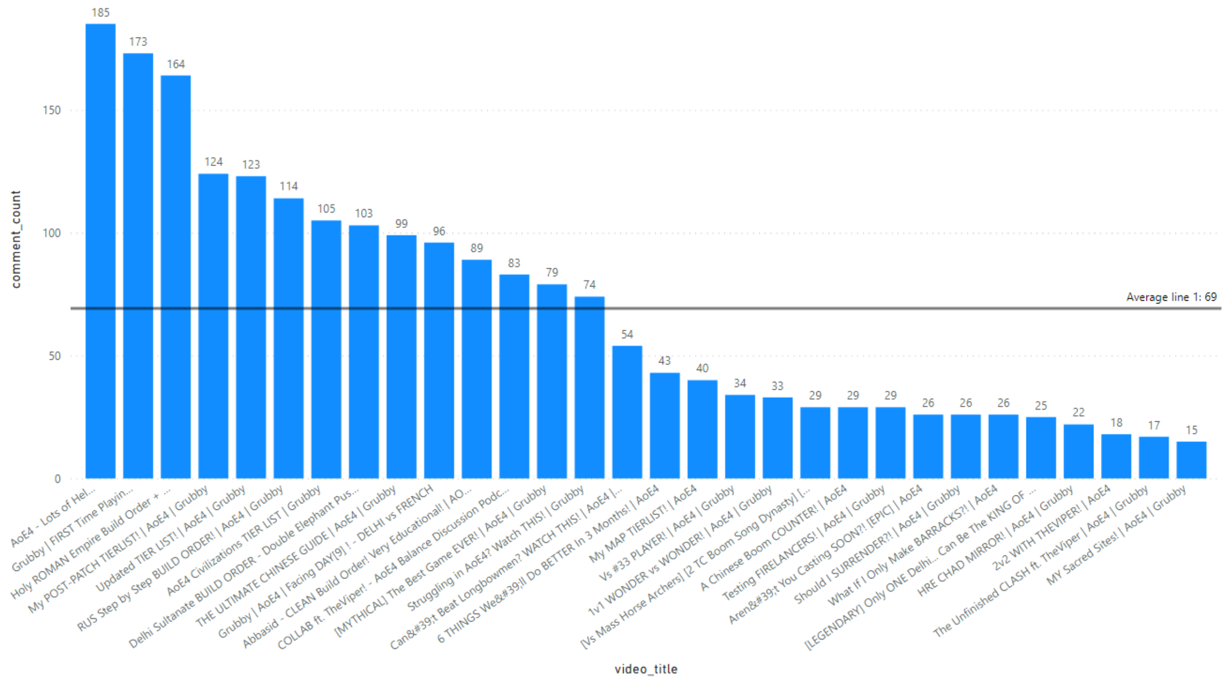


Figure 8 - Number of comments for each vídeo

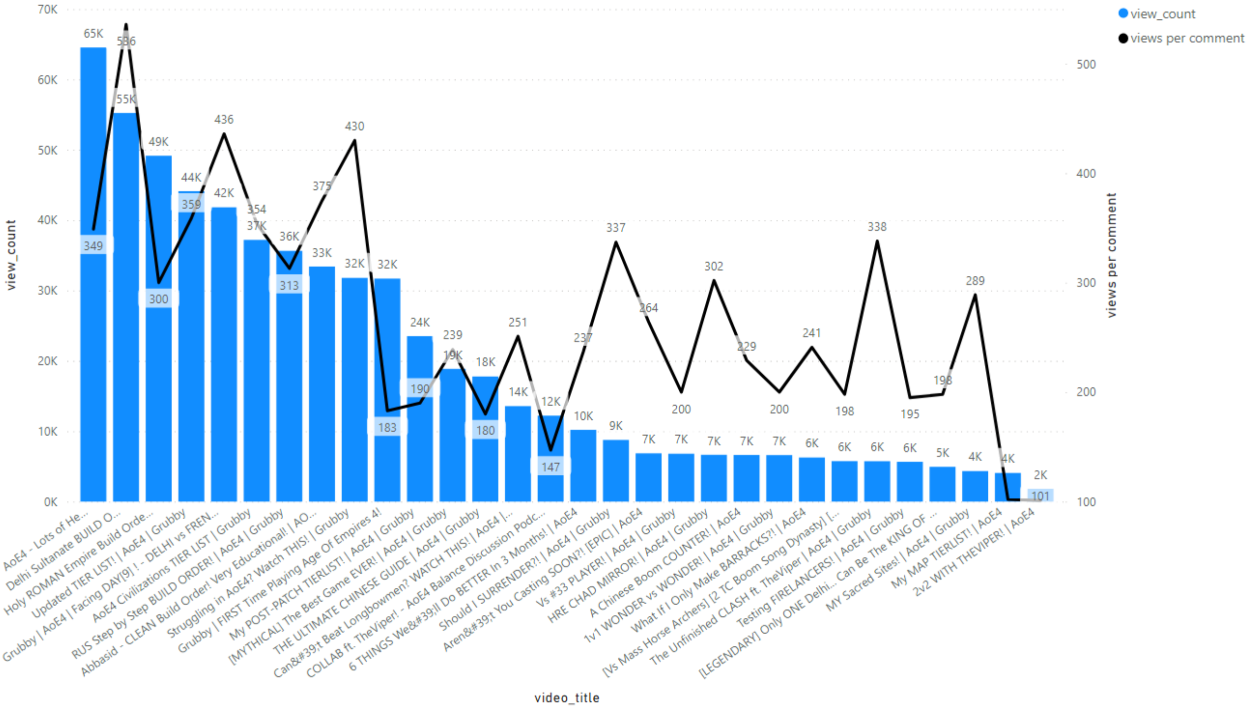


Figure 9 - Comments per view

In the area of comments, we have in figure 8 practically the same result of engagement of likes. Tab videos are always above average comments. In Figure 9, we have how many views are needed to generate 1 comment. An interesting video to highlight is the 'COLLAB ft. TheViper! – AoE4 Balance Discussion Podcast While Watching Replay', because it was the one that had the best result in, perhaps by the podcast $\frac{views}{comments}$format, the only one of all the videos analyzed.

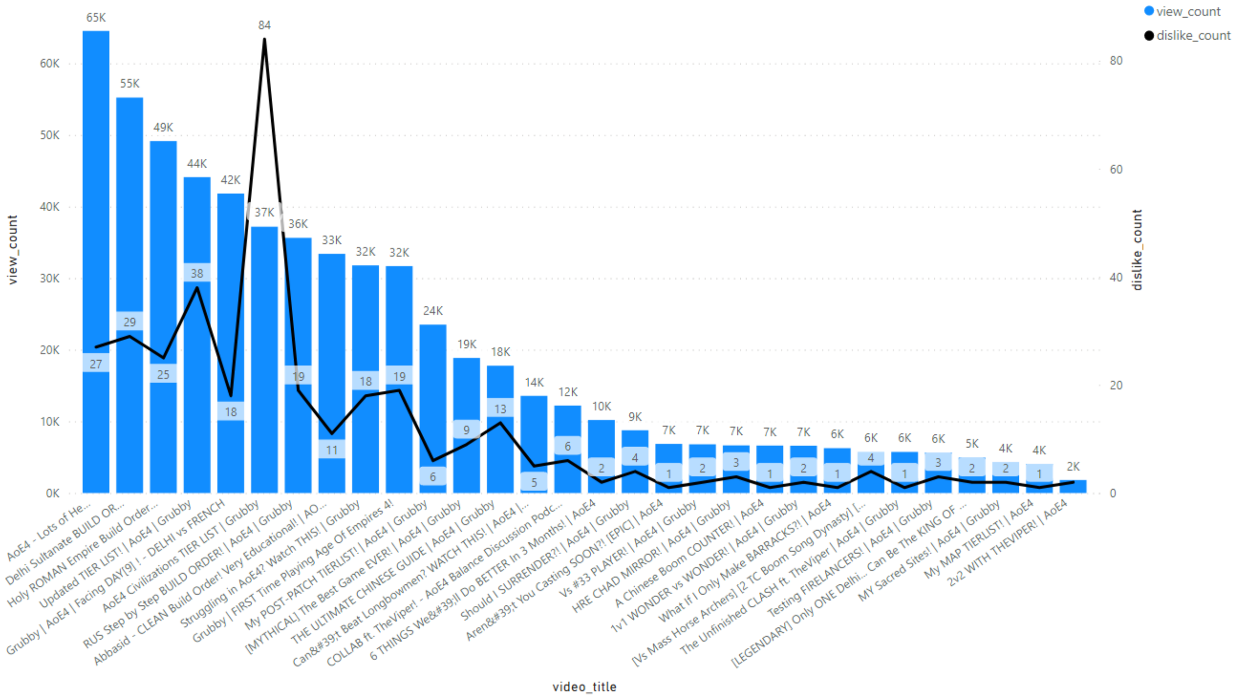


Figure 10 - Dislikes for each vídeo

The channel in general does not have a lot of dislikes, but a video catches the eye. In Figure 10, you can see that the video 'AoE4 Civilizations TIER LIST' received a much larger number of dislikes than the other videos. It is a tierlist video, one of the categories that is making success at the moment, namely a video with personal opinion and apparently people who did not agree gave dislike in the video.

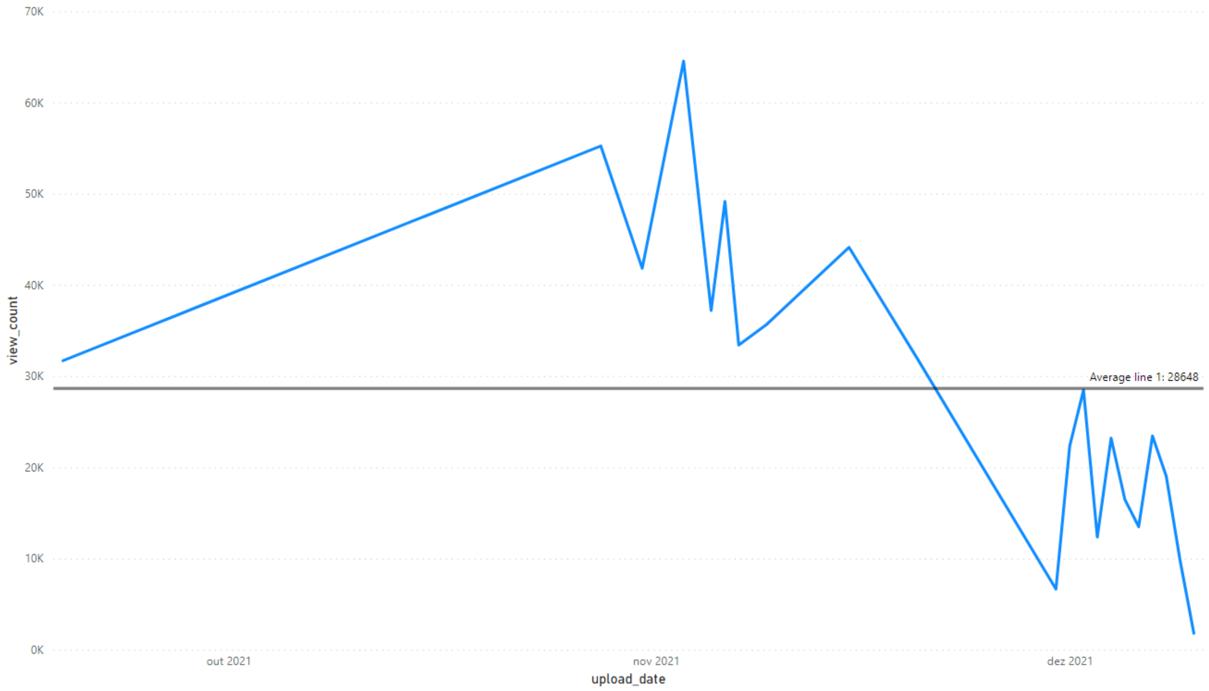


Figure 11 - Number of views for the last quarter of the year

Regarding the number of views in the last quarter of the year, we have in figure 11, that in October (almost no uploads) and November, the number of views were above average and, in December below average.

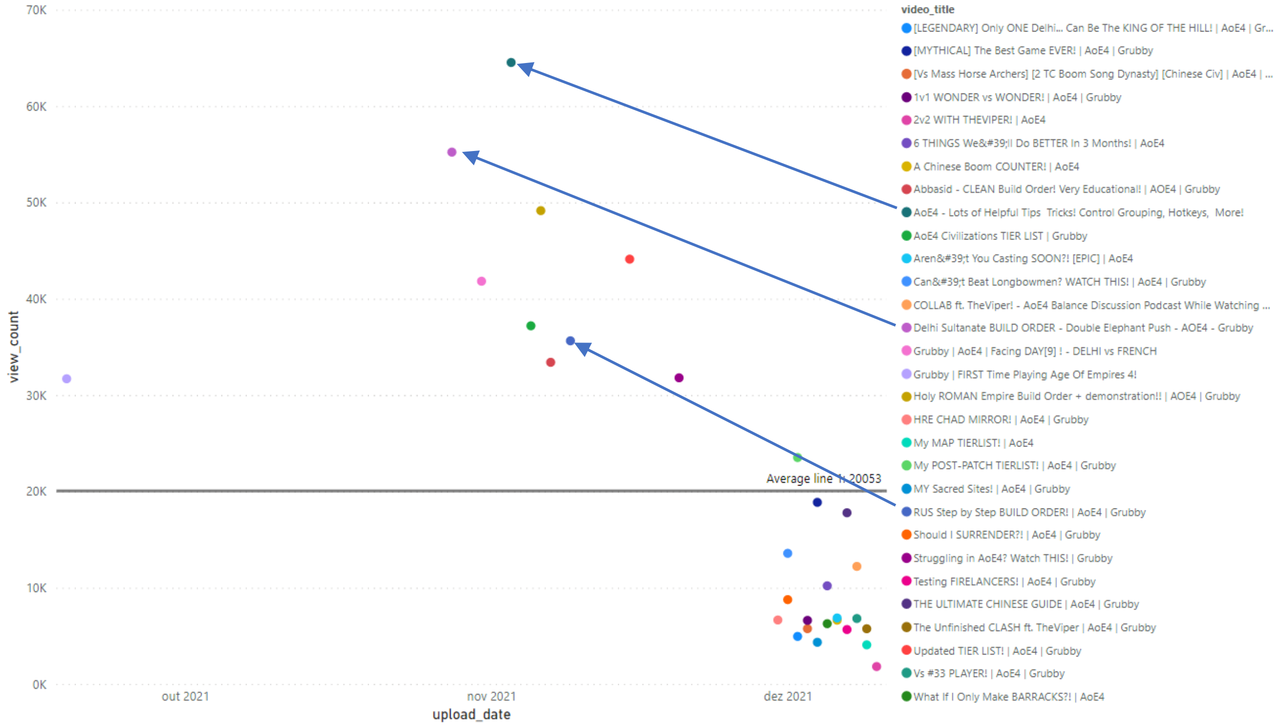


Figure 12 - Number of views for the last quarter of the year with video title

The reason for November is always above the average views and the month of December below is in figure 12. In it we see that the videos of guides and help were all uploaded in the month of November, while the rest in the following month.

1. Conclusions

It was possible to collect the data from the youtube API and store it in a database in the cloud. The analysis showed us that videos with the themes of guides are the most popular and the ones that produce the most community engagement. Finally, we have a video containing personal opinions, in which generated the largest number of dislikes among all the videos analyzed