INSTITUTO POLITÉCNICO DE BEJA

Escola Superior de Tecnologia e Gestão

Licenciatura em Engenharia Informática

Sistemas de Informação

Trabalho Prático nº 1

Elaborado por:

José Francisco Fernandes

Tierri Ferreira

Docentes:

Isabel Brito

13/11/2023

Índice

Figuras

1. Introdução

Neste trabalho temos como objetivo extrair, transformar e carregar dados sobre um tema escolhido por nós para fazer uma análise OLAP aos mesmos. Será também feito um modelo de data warehouse para estes dados.

O tema escolhido está relacionado com dados de vídeos nas tendências do YouTube, filtrados com base nos canais com mais subscritores.

Para fazer o ETL, foi feito um *script* em Python para filtrar os dados com base num dataset dos canais mais subscritos no YouTube para reduzir a quantidade de informação e manter os dados estatísticos nos canais maiores, para evitar excesso de dados que apenas existem devido a uma certa tendência do momento.

O presente relatório encontra-se organizado na seguinte forma:

- Na secção 2 fazemos uma breve descrição dos datasets obtidos;
- Na secção 3 é descrito o processo ETL;
- Na secção 4 demonstra-se o modelo multidimensional (Data Warehouse);
- Na secção 5 é feita a análise OLAP.

2. Datasets obtidos

Os datasets foram obtidos a partir do Kaggle [1] [2] e estão relacionados com dados de vídeos nas tendências no YouTube e os canais mais subscritos do YouTube, respetivamente.

Ambos estes datasets estão formatados em CSV, mas o primeiro dataset (com os vídeos nas tendências) também tem um ficheiro JSON de categorias à parte com os dados das categorias dos vídeos com o seu nome, em vez de um ID numérico. Este ficheiro foi também usado para fazer a análise OLAP.

Adicionalmente, o dataset de vídeos nas tendências traz diversos ficheiros CSV com países diferentes, uma vez que as tendências variam consoante a localização geográfica. Neste trabalho, escolhemos os EUA.

O segundo dataset (canais mais subscritos no YouTube) foi usado para manter a análise dos dados mais abrangente, ou seja, uma vez que estamos a trabalhar ao nível estratégico, não queremos que tendências atuais influenciem dados. Por exemplo, se houver uma tendência que dura um mês e acontece num canal pouco famoso, este(s) vídeo(s) influenciarão os resultados da análise OLAP e os dados poderão ser menos realistas.

Para tornar esta filtragem possível, usámos um *script* Python que filtrava os vídeos nas tendências pelos canais mais subscritos. Não foi possível fazer esta filtragem sem este dataset devido ao facto de dados de subscritores apenas existirem em datasets de canais, e não em datasets de vídeos. Na secção 3 será possível visualizar este processo.

3. Processo ETL

Para tornar o processo ETL possível, foram usadas duas tecnologias principais. Inicialmente, para o primeiro *extract*, foi usado o Python com a biblioteca *csv*. Na Figura 1 é possível observar este *script* simples.

```python
import csv
import json

def parseCsvToArray(filename):
    reader = csv.reader(
        open(filename, 'r', encoding='utf-8'))
    data = []
    headers = next(reader)
    nextLine = next(reader)
    while nextLine != None:
        obj = {}
        for i in range(len(headers)):
            obj[headers[i]] = nextLine[i]
        data.append(obj)
        try:
            nextLine = next(reader)
        except StopIteration:
            break
        except Exception:
            continue

    return data

videos = parseCsvToArray('US_youtube_trending_data.csv')

years = []
for i in range(len(videos)):
    year = videos[i]['publishedAt']
    sub = year[0:4]
    if sub not in years:
        years.append(sub)

channels = parseCsvToArray('most_subscribed_youtube_channels.csv')

filteredVideos = []
for i in range(len(channels)):
    channelVideos = list(filter(lambda c: c['channelTitle'] == channels[i]['Youtuber'], videos))
    filteredVideos = filteredVideos + channelVideos

with open('out.json', 'w') as f:
    json.dump(filteredVideos, f)

print('success')
```

Figura 1 – *Script* em Python para o primeiro *extract* e *transform*.

Na função definida na linha 4, *parseCsvToArray*, é feita a leitura do ficheiro CSV e já é feita uma transformação muito simples, que é a eliminação de dados corruptos/incorretos na sintaxe CSV. Isto foi observável no maior dataset, constituído pelos vídeos nas tendências.

Nas linhas 26 a 31, apenas foi feita uma verificação do espaço de tempo dos dados, para que se verificasse se este dataset tinha um intervalo de tempo suficientemente grande.

Seguidamente é feita a filtragem a partir do segundo dataset, que é lido da mesma forma que o primeiro. Estes dados são iterados (os canais) e faz-se assim uma filtragem dos vídeos, de modo que apenas vídeos do canal iterado sejam adicionados ao array final.

Este array final, por fim, é enviado para um ficheiro JSON de saída, que depois será interpretado pelo PowerBI.

Para a segunda e última parte do *extract* e *transform*, foi usado o PowerBI para alterações mais simples.

A primeira coisa que foi feita foi a abertura de um novo projeto no PowerBI e o carregamento do ficheiro JSON *out.json*. Isto abrirá o Power Query, como se observa na Figura 2.

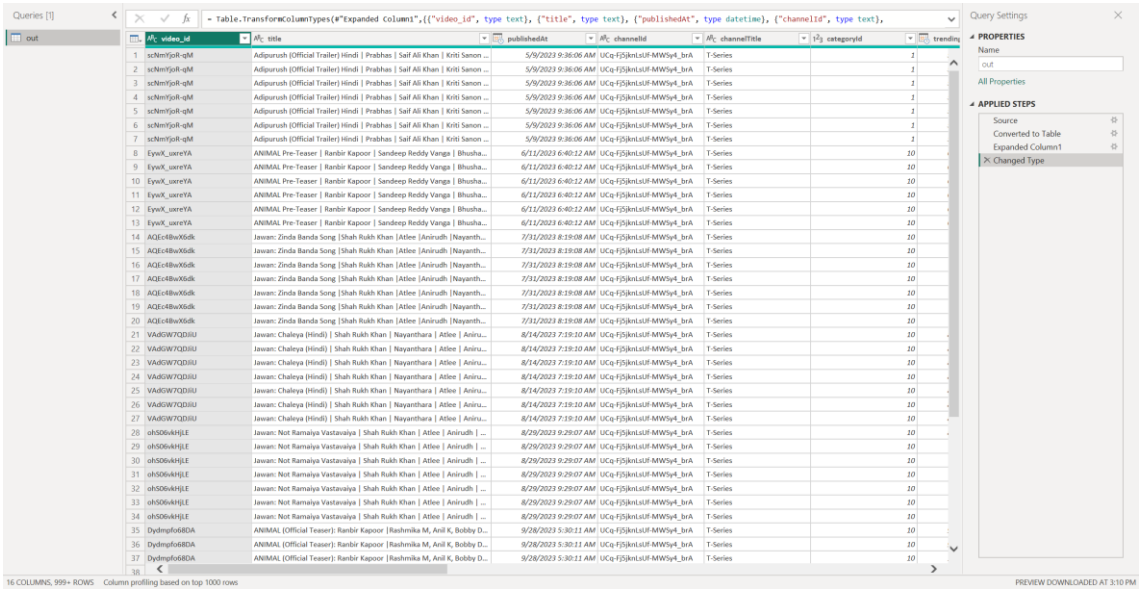

Figura 2 – Power Query, aberto ao carregar o ficheiro out.json no PowerBI, com os vídeos filtrados no *script* Python.

Por agora, não será necessário editar nada. Simplesmente fecharemos e carregaremos estes dados para o PowerBI.

Uma vez que os dados são pouco legíveis devido aos seus nomes, mudaremos os mesmos, como se vê nas Figuras 3 e 4.

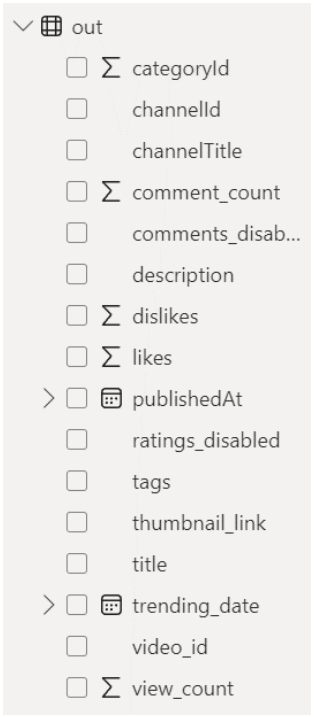

Figura 3 – Nomes anteriores nos dados recentemente carregados.

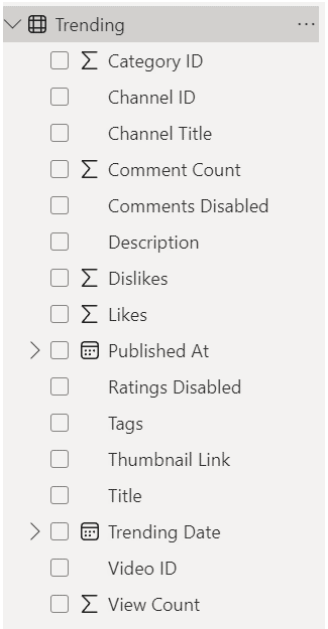

Figura 4 – Nomes atualizados.

Agora, uma vez que na análise vamos precisar de verificar horas de publicação, extraímos as horas de "Published At" para uma nova coluna. Na Figura 5 observamos o código em M para fazer esta coluna.

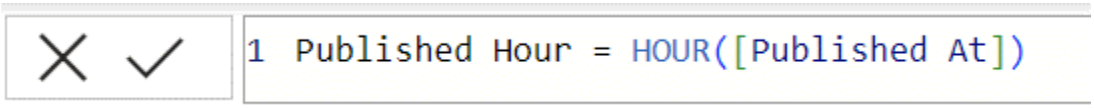

Figura 5 – Criação da coluna "Published Hour".

Foi necessária a criação desta coluna uma vez que o PowerBI, por alguma razão que não foi possível verificar, não permite a filtragem de conteúdo por hora, apenas até ao dia, mesmo que o tipo de dados seja *datetime*.

De seguida, para tornar a leitura mais fácil, vamos alterar tipos booleanos em texto para que seja mais fácil de interpretar a informação analisada. Para fazê-lo, abrimos o Power Query e alteramos os tipos das colunas “Comments Disabled" e “Ratings Disabled”. O tipo é alterado conforme se apresenta na Figura 6.

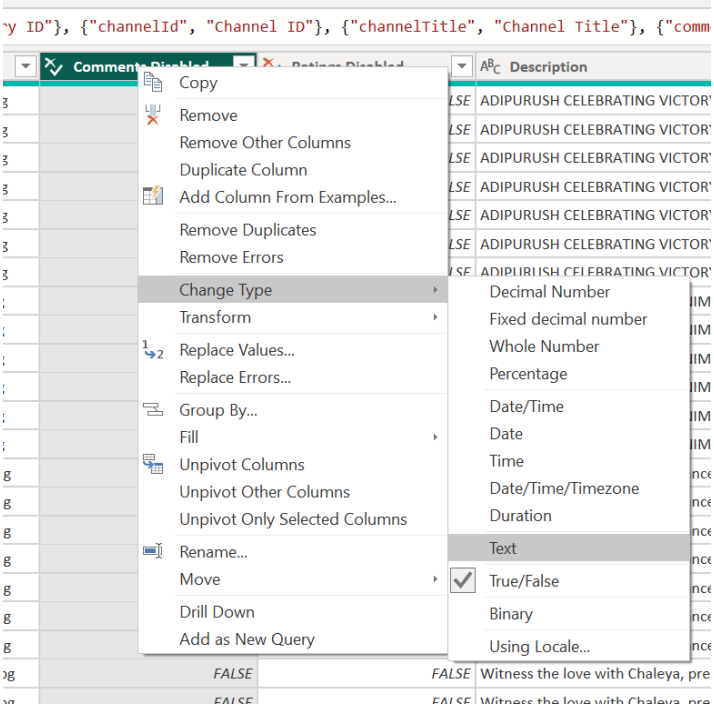

Figura 6 – Alteração do tipo de dados de colunas booleanas.

Após esta alteração, vamos substituir todos os valores “true” e “false” em “sim” e “não”, respetivamente. Na Figura 7 podemos observar como o mesmo se faz.

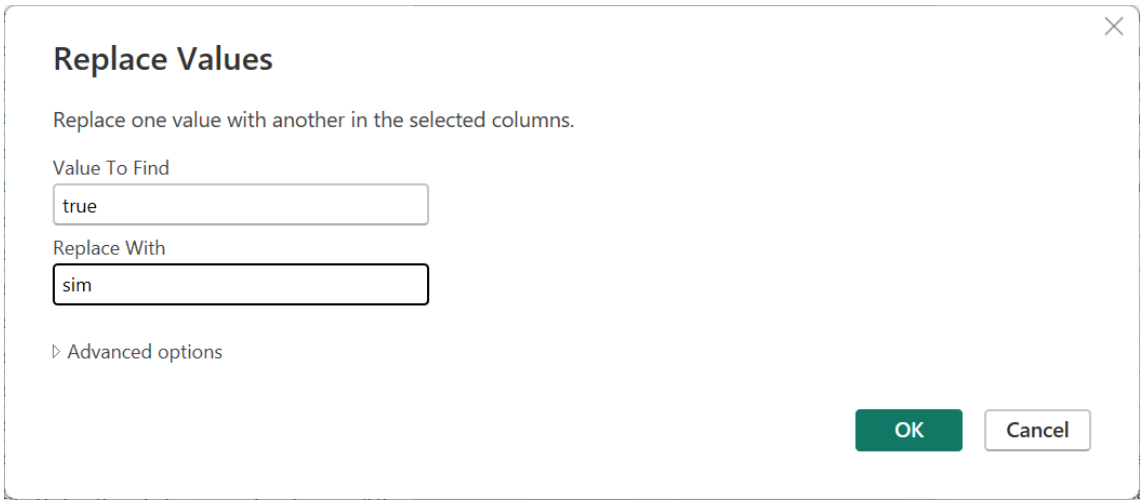

Figura 7 – Substituição de valores booleanos em texto português.

Após esta substituição, podemos novamente carregar os dados do Power Query para o PowerBI.

Agora que terminámos as alterações nos vídeos, vamos carregar as categorias, que se encontram em JSON. Isto novamente abrirá o Power Query.

Desta vez, iremos já fazer a transformação destes dados. Iremos eliminar colunas/atributos redundantes e desnecessárias do ficheiro, uma vez que apenas queremos, principalmente, os nomes das categorias, que não se encontram no dataset principal, para depois fazer a análise com categorias legíveis.

Eliminaremos as seguintes colunas/atributos:

- kind (atributo)
- etag (atributo)
- items.kind (coluna)
- items.snippet.assignable (coluna)
- items.snippet.channelId (coluna)
- items.etag (coluna)

Os únicos dados que restam desta transformação são os seguintes (Figura 8):

	items.id	items.snippet.title
1	1	Film & Animation
2	2	Autos & Vehicles
3	10	Music
4	15	Pets & Animals
5	17	Sports
6	18	Short Movies
7	19	Travel & Events
8	20	Gaming
9	21	Videoblogging
10	22	People & Blogs
11	23	Comedy
12	24	Entertainment
13	25	News & Politics
14	26	Howto & Style
15	27	Education
16	28	Science & Technology
17	29	Nonprofits & Activism
18	30	Movies
19	31	Anime/Animation
20	32	Action/Adventure
21	33	Classics
22	34	Comedy
23	35	Documentary
24	36	Drama
25	37	Family
26	38	Foreign
27	39	Horror
28	40	Sci-Fi/Fantasy
29	41	Thriller
30	42	Shorts
31	43	Shows
32	44	Trailers

Figura 8 – Categorias de vídeos.

Os nomes das colunas e do ficheiro foram alterados, tal como no anterior. Na Figura 9 observam-se os novos nomes.

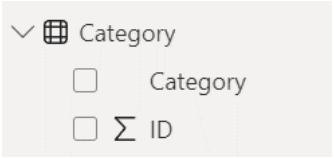

Figura 9 – Novos nomes das colunas de Category.

Por fim, foi feita uma relação entre Trending e Category para que o ID das categorias fosse reconhecido como chave estrangeira e o nome das categorias seja diretamente associado durante a fase de análise.

Na Figura 10 podemos observar esta relação.

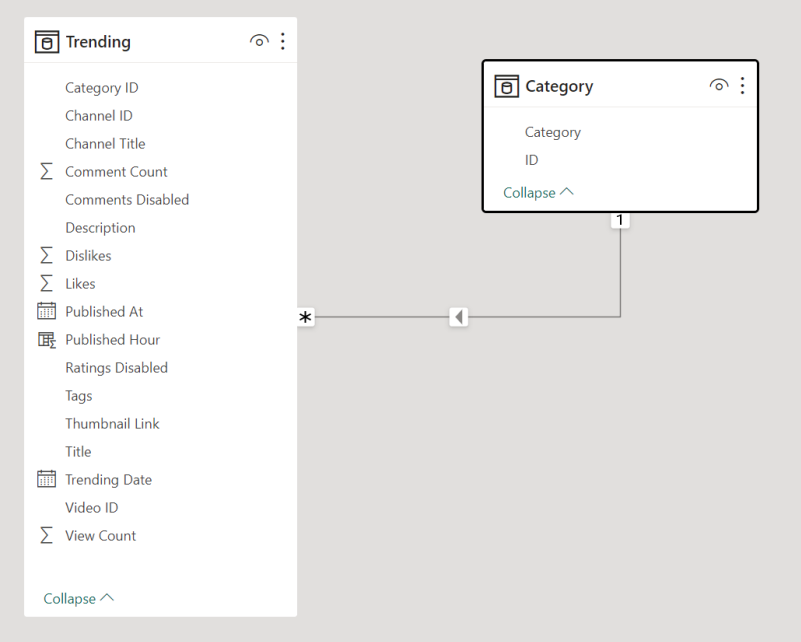

Figura 10 – Relação entre Trending e Category.

Para criar esta relação, simplesmente arrasta-se o atributo “Category ID” em Trending para “ID” em Category e o PowerBI automaticamente cria a relação.

4. Modelo Multidimensional

Com o processo de ETL feito foi necessário definir um modelo multidimensional, para ser mais fácil identificar o que queremos estudar com estes dados.

Optámos por fazer uso do modelo em estrela, começando por definir a tabela de facto que será o centro. A tabela chama-se FactTrending.

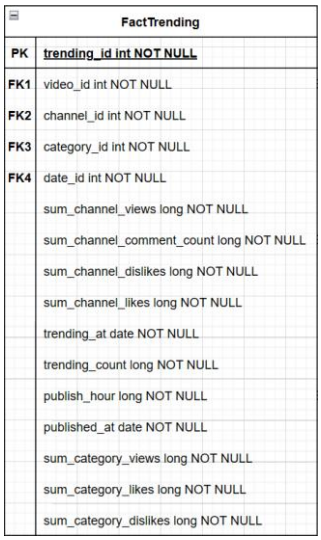

Figura 11 – Tabela de facto (FactTrending)

Mas para definir a tabela de facto é necessário construir as tabelas de dimensão. Estas são quatro, uma delas sendo a tabela DimDate que é obrigatória nos modelos multidimensionais.

As restantes podem ser observadas na Figura 12.

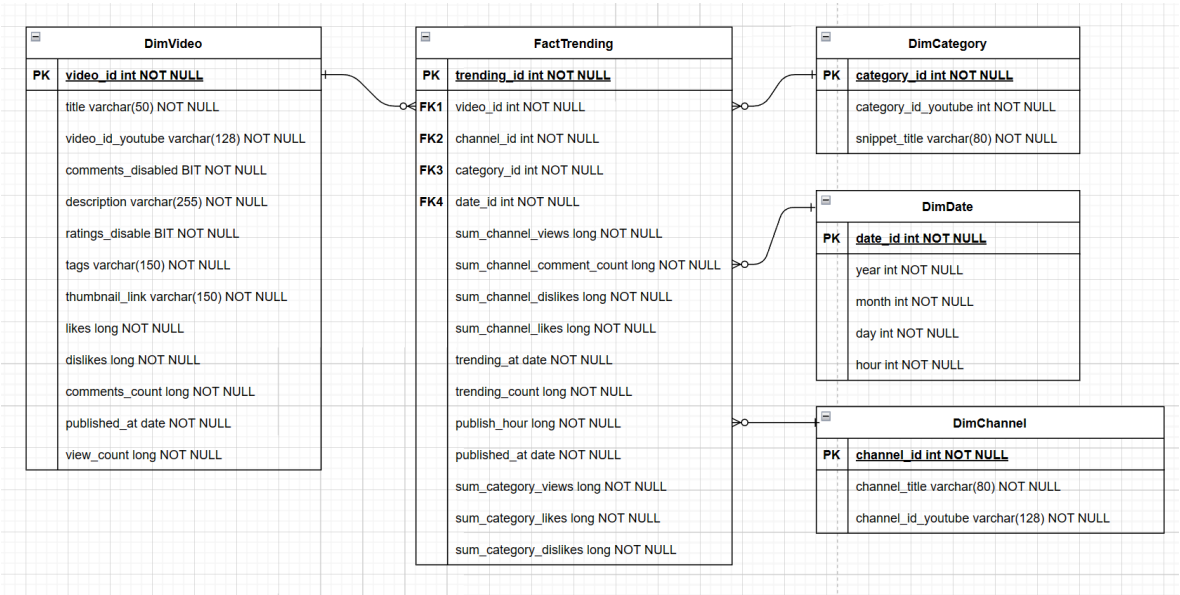

Figura 12 – Modelo multidimensional.

A tabela DimVideo, tal como o nome indica, representa um vídeo do YouTube. Seguidamente temos também a DimChannel que representa os canais e a DimCategory que contém as categorias de vídeos nos EUA.

Escolhemos estes dados na tabela de facto através de uma análise prévia aos dados do dataset para que se observasse o que poderia ser analisado. Tirámos a conclusão de que os valores principais são as visualizações, os gostos/não-gostos e o número de comentários, em diversos somatórios relacionados com o canal ou com uma categoria, por exemplo.

Na análise OLAP, todos estes campos são usados eventualmente na realização dos gráficos e na visualização da análise deste dataset.

5. Análise OLAP

Para a análise OLAP foram usados os dados já extraídos, transformados e carregados no PowerBI.

Esta análise acontece no intervalo de tempo dos anos 2020 a 2023 (atualidade, uma vez que o dataset usado é atualizado ao dia).

Na Figura 13 é feita uma comparação entre quantidade de visualizações e gostos em vídeos por ano e mês.

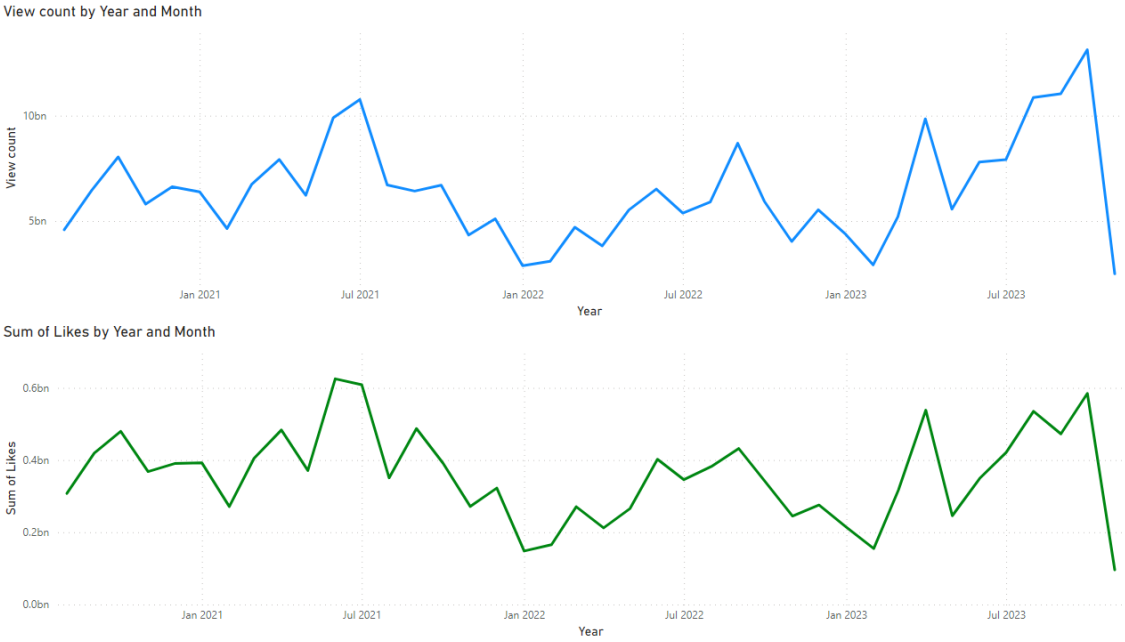

Figura 13 – Contagem de visualizações/gostos por ano e mês.

Com este gráfico, pode-se concluir que existe relação entre a quantidade de gostos e a quantidade de visualizações devido à semelhança visual entre os mesmos. É também possível verificar que a quantidade de visualizações/gostos (ou seja, qualquer atividade no YouTube) é mais visível no Verão, devido ao facto de esta ser uma época de férias.

Na Figura 14 é feita uma visualização da quantidade de visualizações de vídeos nas tendências que têm os comentários ativados, comparado com vídeos com comentários desativados.

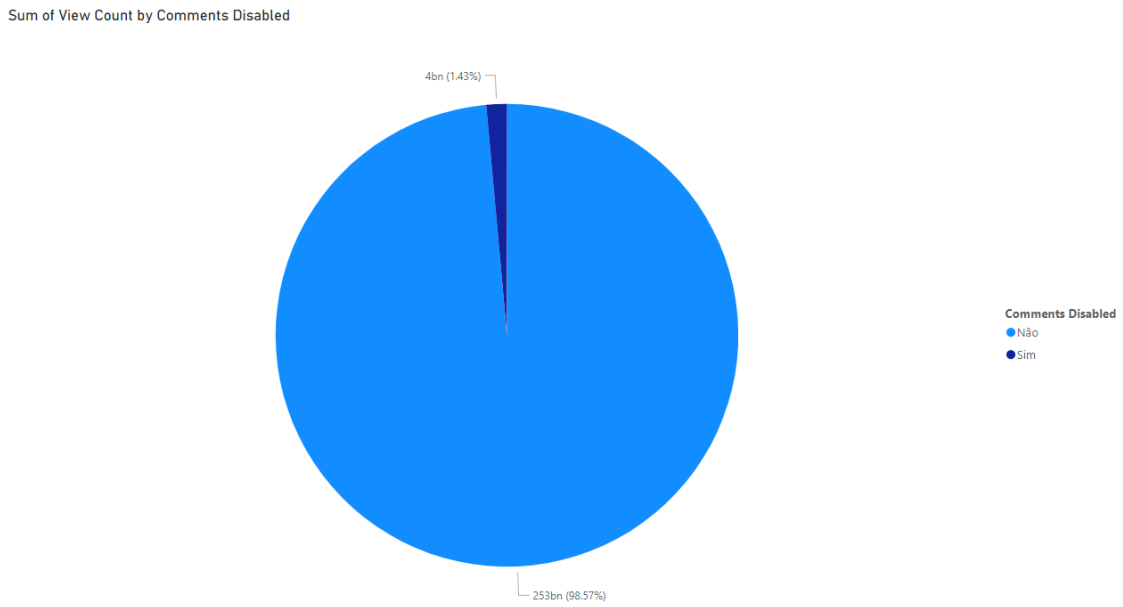

Figura 14 – Visualizações em vídeos com comentários ativados vs. vídeos com comentários desativados.

Pode-se concluir que, na grande maioria (98.57%) dos vídeos nas tendências com os comentários ativados existem muito mais visualizações do que em vídeos com comentários desativados. No entanto, isto é também devido ao facto de ser raro os canais desativarem esta funcionalidade, tal como se observa na situação dos *ratings* (quantidade de gostos e não-gostos públicos). O gráfico para estes dados encontra-se na Figura 15.

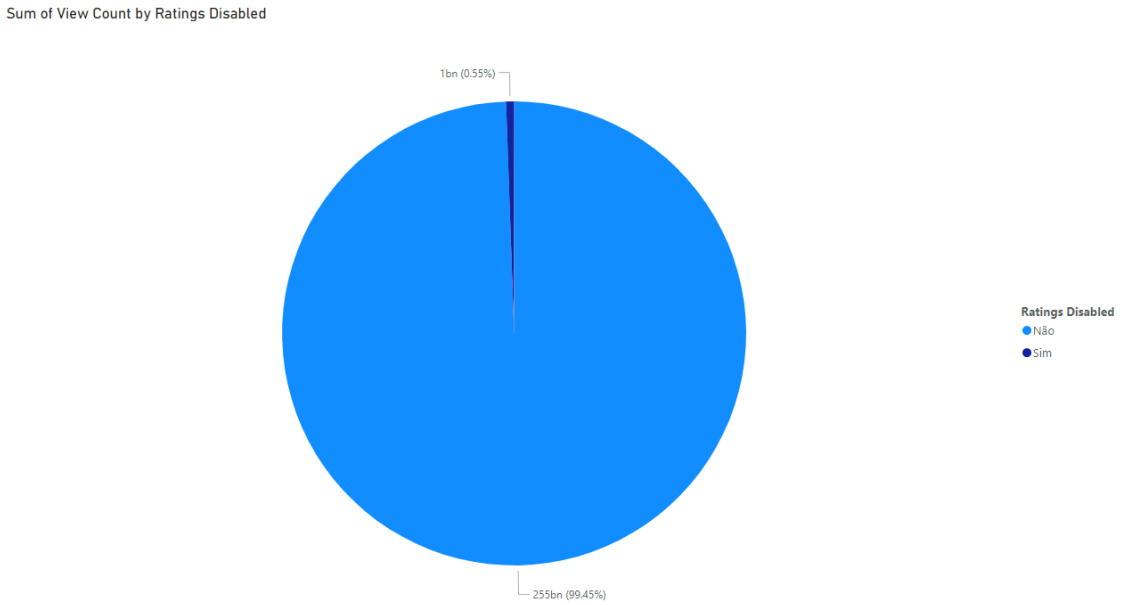

Figura 15 – Visualizações em vídeos com *ratings* ativados vs. vídeos com *ratings* desativados.

De seguida, foi feita uma análise à hora de publicação de vídeos. Isto porque, muitos canais de grande dimensão no YouTube encontram horas mais favoráveis para fazer a publicação dos seus novos vídeos. Na Figura 16 podemos observar os resultados da análise feita, através de uma visualização de gostos e visualizações de vídeos publicados em determinada hora.

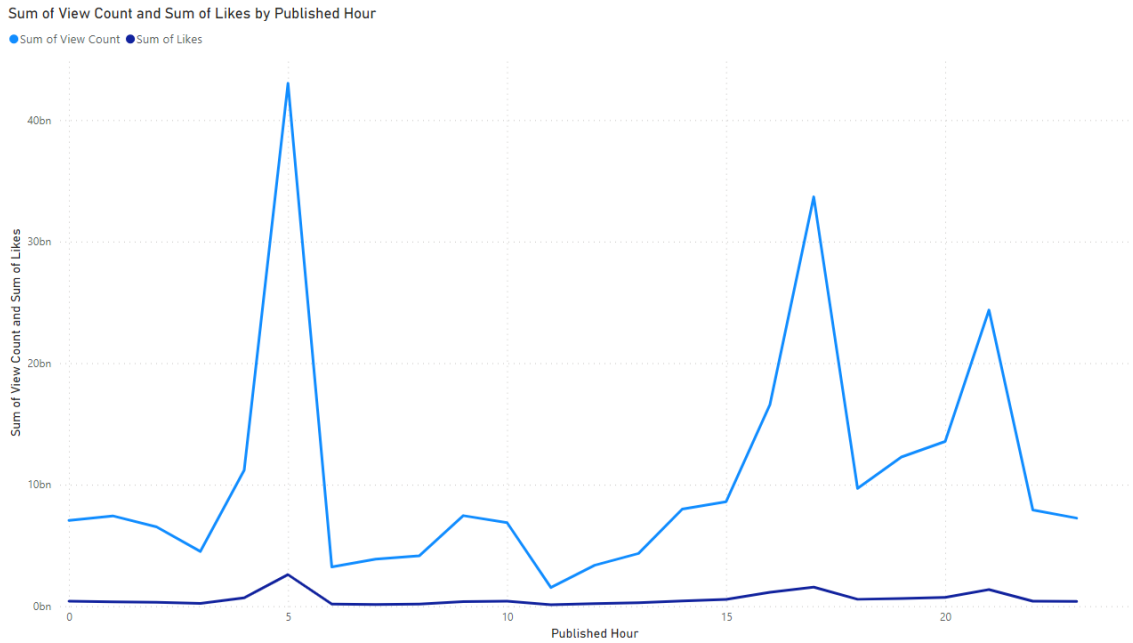

Figura 16 – Hora de publicação de vídeos e a sua quantidade de visualizações e gostos.

Podemos tirar a simples conclusão de que, nos Estados Unidos, há melhores resultados ao publicar vídeos às 5 horas da manhã (sendo a melhor hora), às 5 horas da tarde e às 9 horas da noite.

Após esta análise, foram feitos alguns gráficos para verificar quais são as categorias que têm mais atividade/encontram-se mais nas tendências.

Na Figura 17 é feita uma relação entre gostos/não-gostos nas categorias mais populares.

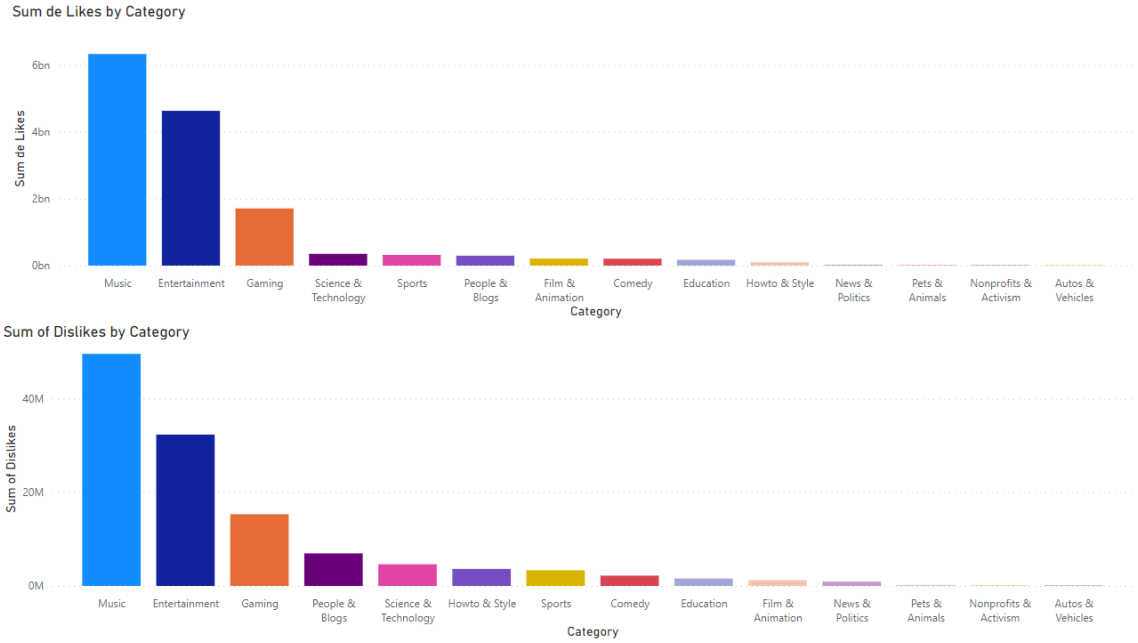

Figura 17 – Relação entre gostos e não-gostos em categorias.

Uma vez que esta plataforma é principalmente de entretenimento, é evidente que uma grande parte da sua atividade se encontre na categoria de entretenimento. Adicionalmente, a música sempre foi (e cresceu ainda mais como categoria com o lançamento do YouTube Music) uma das categorias mais famosas do YouTube devido ao facto de este ser usado para muitas vezes partilhar músicas.

Existe pouca diferença entre gostos e desgostos, mas ainda pode-se observar alguma diferença nas categorias menos populares.

Uma análise mais geral foi feita na Figura 18, em que se faz uma análise às categorias, mas em total de visualizações.

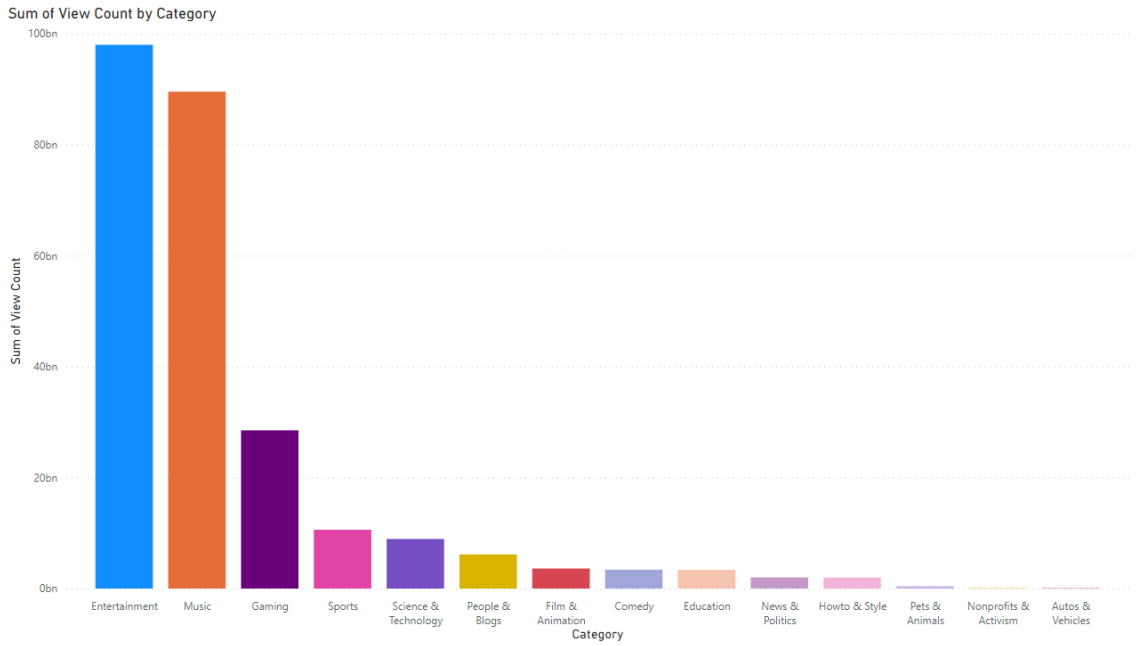

Figura 18 – Número de visualizações por categoria.

Os números são muito semelhantes aos do gráfico anterior, mas vemos uma pequena diferença entre o entretenimento e a música. Pode-se justificar esta diferença devido ao elevado número de visualizações de vídeos de entretenimento, mas ao elevado número de gostos/não-gostos em música devido ao YouTube Music, que é muito usado por qualquer utilizador dessa aplicação quando gosta/desgosta de certa música que acabou de ouvir na plataforma.

De seguida, foi feita uma breve análise ao número de tendências do YouTube por ano, para fazer uma rápida análise à atividade do YouTube por ano. Note-se que uma tendência não é necessariamente única para um vídeo em específico: Uma tendência é criada quando atividade aumenta num específico vídeo num curto intervalo de tempo.

Podemos observar esta análise simples na Figura 19.

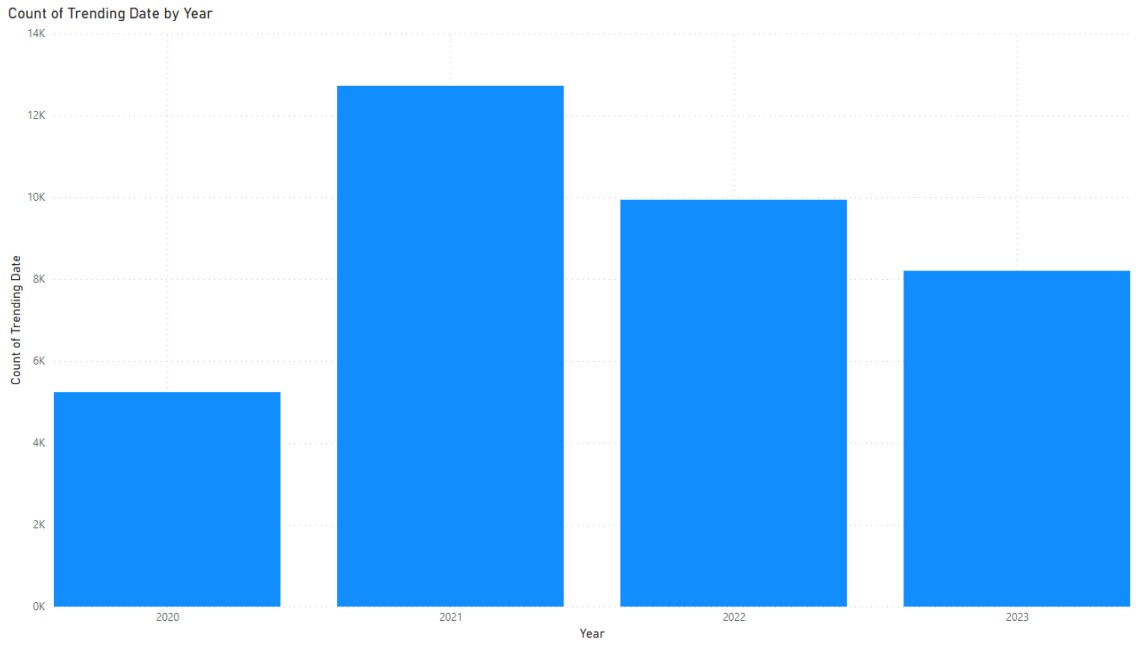

Figura 19 – Número de tendências por ano.

No ano de 2021, vemos um grande aumento de atividade (mais do dobro) comparado com o ano anterior. Não podendo tirar conclusões concretas, podemos assumir que a pandemia teve um forte impacto nesta subida.

Depois desta análise, fizemos algumas análises por canal individualmente, assim como a vídeos individuais. Na Figura 20, são listados os canais com mais comentários nas tendências deste dataset.

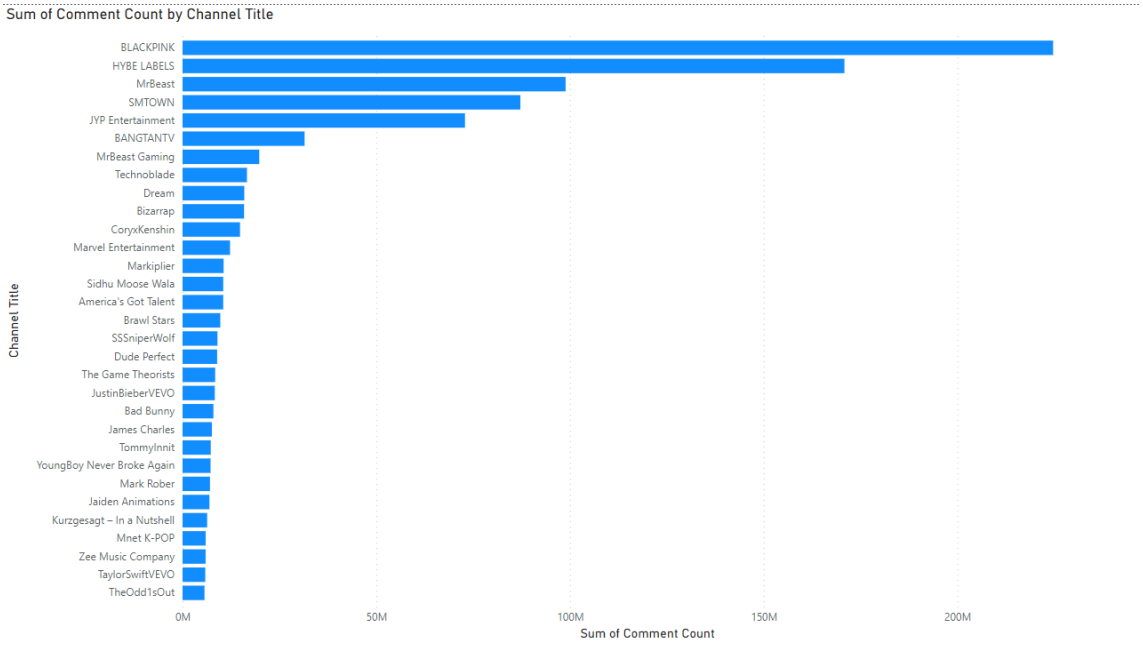

Figura 20 – Número de comentários por canal.

Na Figura 21, faz-se uma análise semelhante, mas por vídeo. São também visíveis as categorias de cada vídeo.

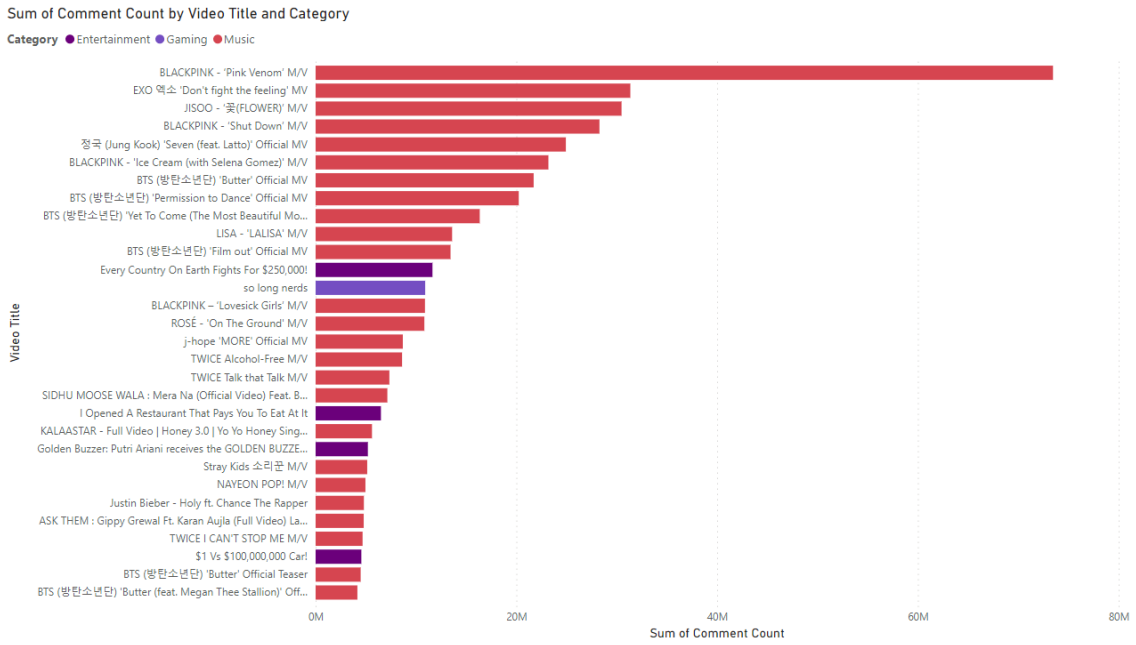

Figura 21 – Número de comentários por vídeo, com categoria.

Pode-se observar uma grande tendência na música, novamente, mas de forma muito mais acentuada. Apenas existem quatro vídeos de entretenimento e um de *gaming*. Uma vez que estes dados são sobre comentários, existe mais diferença nos dados comparado a gostos ou visualizações. Isto porque utilizadores tendem a comentar mais em certas categorias.

Na Figura 22 é feita uma análise semelhante, mas por número de visualizações.

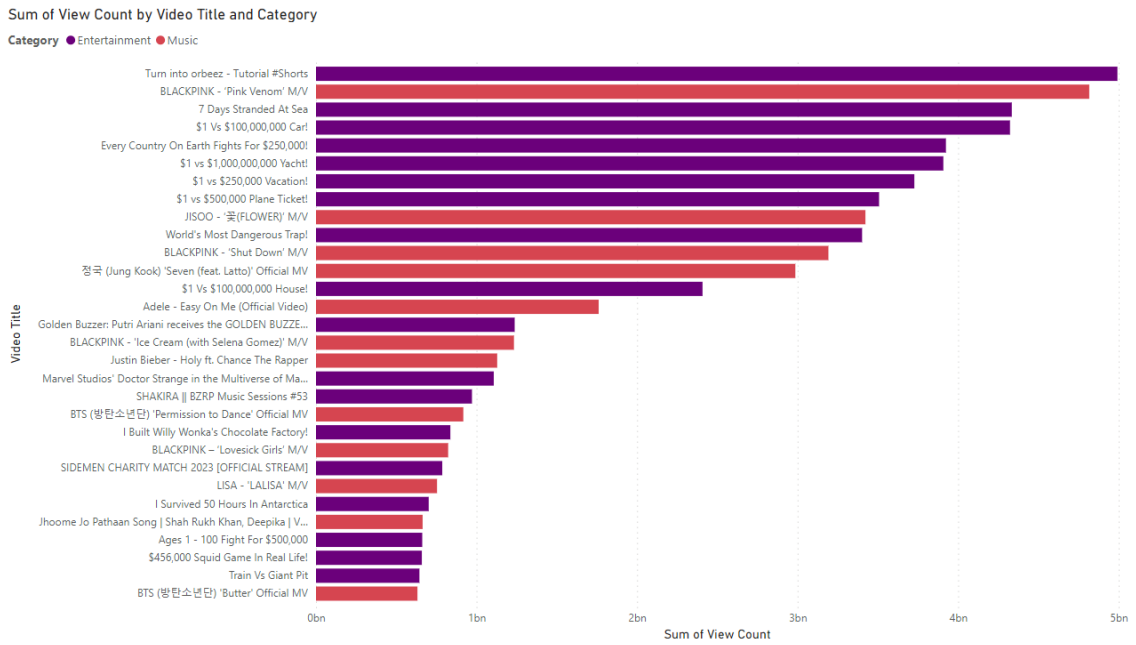

Figura 22 – Número de visualizações por vídeo, com categoria.

Podemos observar um número de visualizações muito mais alto no entretenimento. Como mencionado no gráfico anterior, o número de comentários tem resultados diferentes em relação ao número de visualizações, devido às tendências de comentar/visualizar diferentes entre categorias e canais.

Na Figura 23 é feita uma análise novamente por canal, mas pelas suas visualizações.

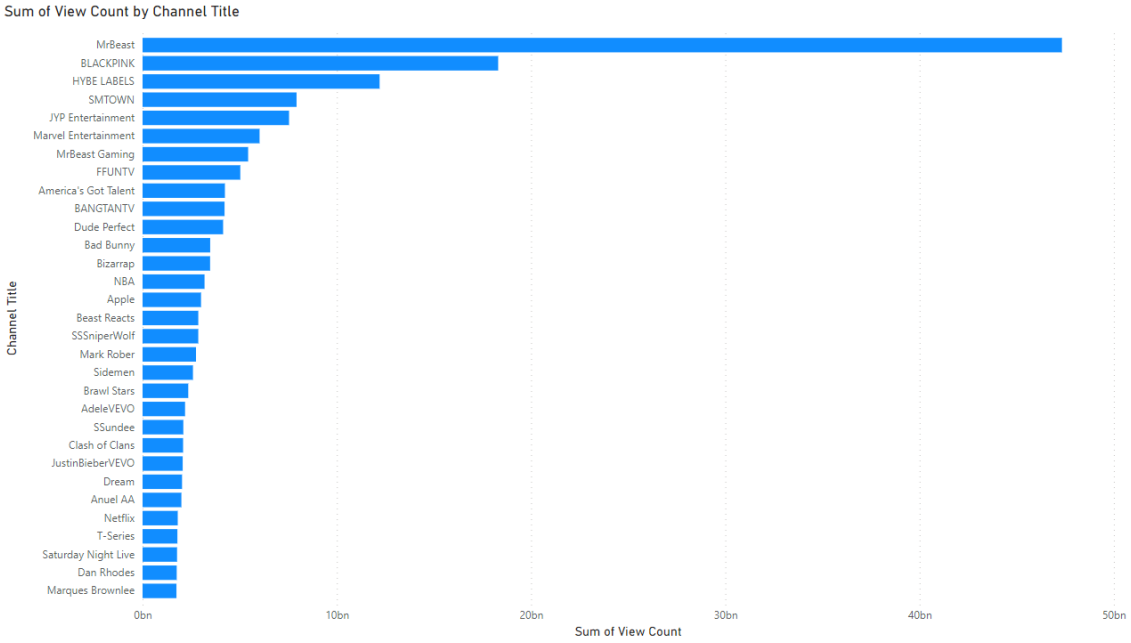

Figura 23 – Número de visualizações por canal.

Podemos ver um número de visualizações extremamente alto no canal *MrBeast*, que é maioritariamente de entretenimento.

Na Figura 24 verifica-se o número de não-gostos em cada canal.

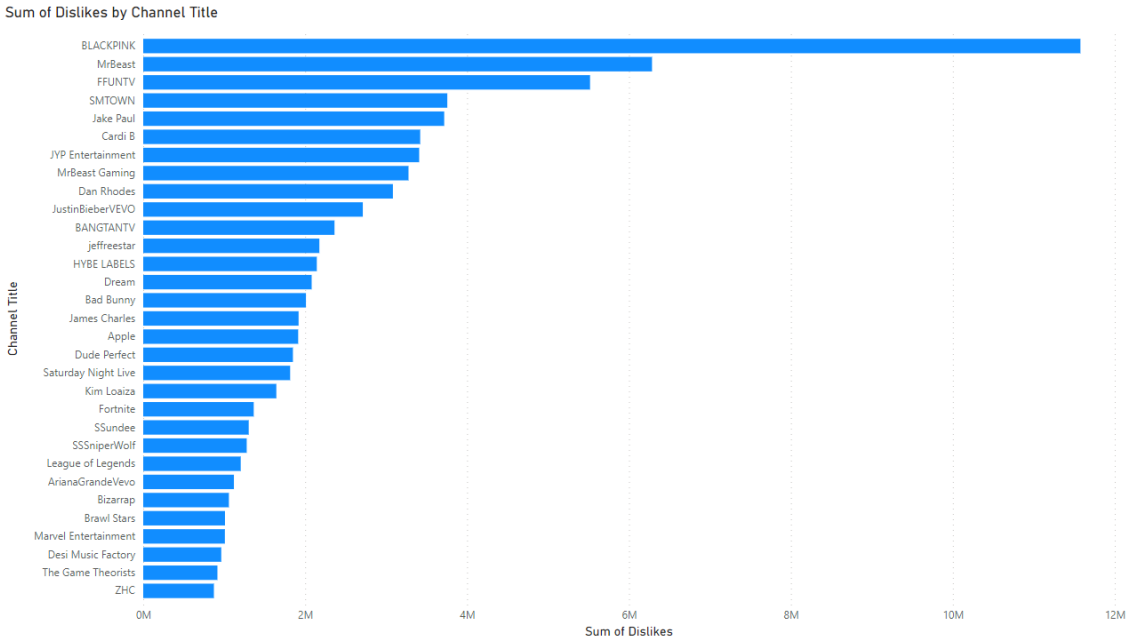

Figura 24 – Número de não-gostos por canal.

Podemos verificar uma relação entre os não-gostos e o número de visualizações, podendo observar que estes nem sempre estão relacionados com a qualidade do vídeo, mas sim com a popularidade do canal.

Agora, é feita uma análise ao número de vezes que um canal esteve nas tendências, na Figura 26.

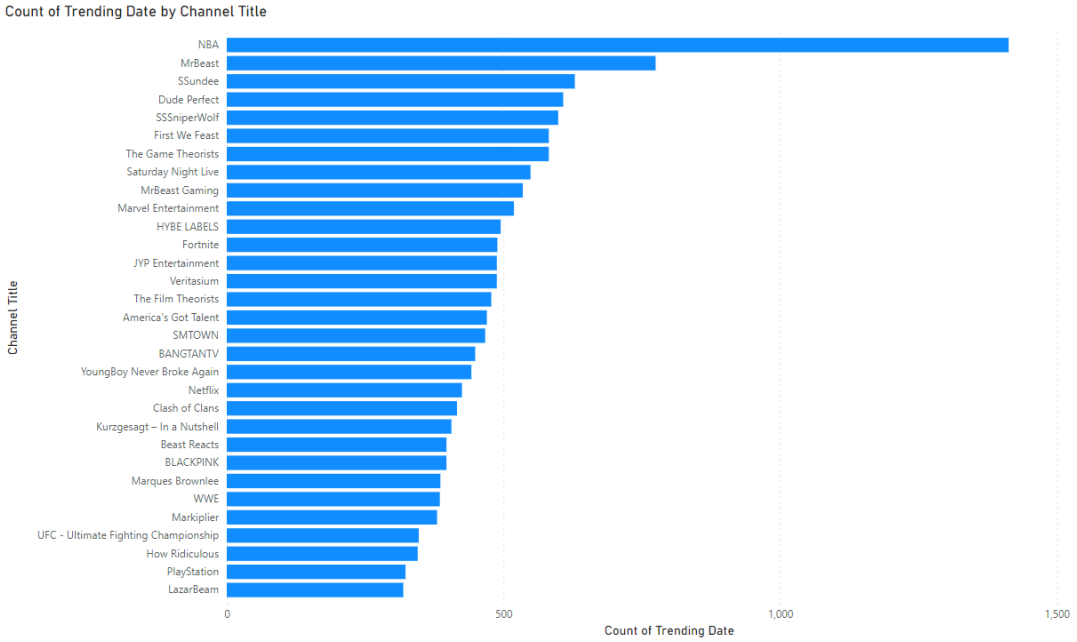

Figura 26 – Número de vezes que cada canal esteve nas tendências.

Podemos observar que o canal da NBA apareceu neste gráfico quando nem estava no gráfico anterior. Isto pode ser devido à sua categoria. Uma vez que é um canal de desporto, há mais probabilidades de haver uma tendência após, por exemplo, um jogo, uma vez que há influência de dados externos neste canal, enquanto a tendência de outros canais acontece devido a tendências do conteúdo do próprio canal.

Outra razão poderá ser o número elevado de publicações deste canal, como se pode verificar no gráfico da Figura 27.

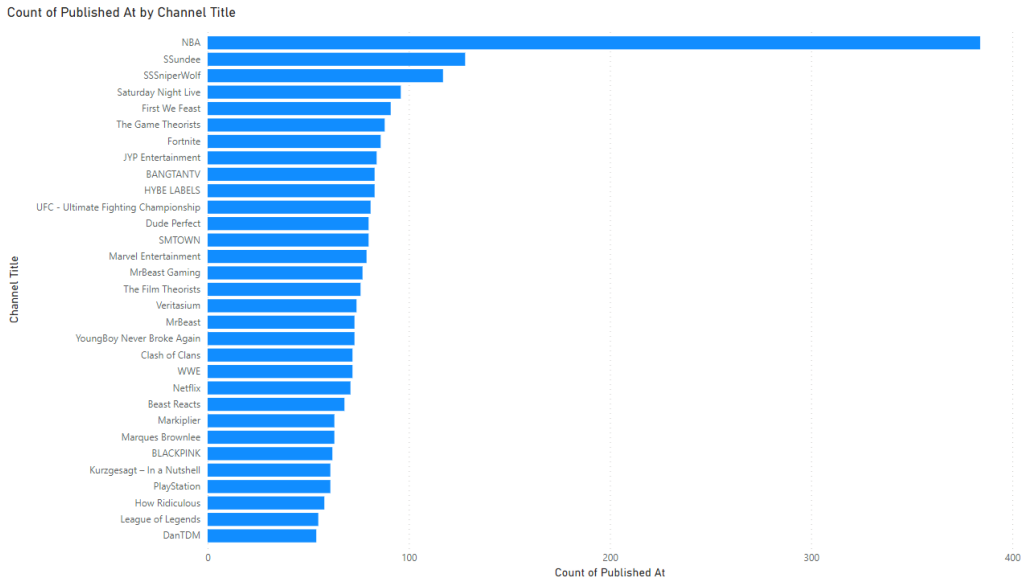

Figura 27 – Número de publicações por canal.

Na Figura 28 é feita uma análise aos vídeos que entraram mais nas tendências no intervalo de tempo deste dataset, com a respetiva categoria.

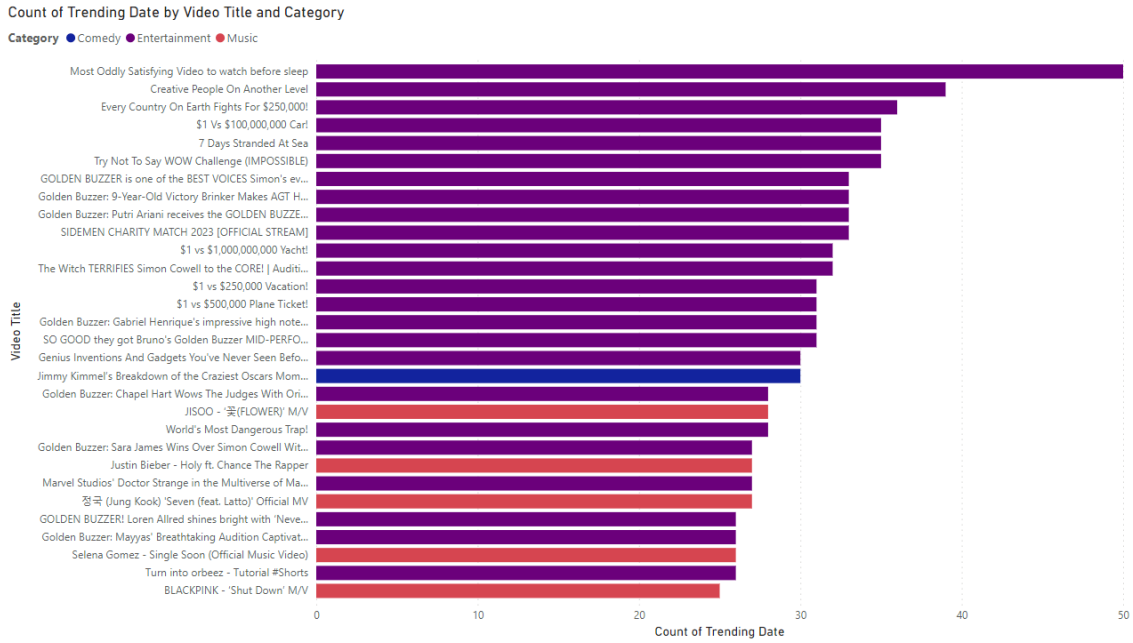

Figura 28 – Número de vezes que vídeos entram nas tendências com a categoria.

Mais uma vez, os vídeos que entram mais nas tendências são quase sempre sobre entretenimento e música. O único dado menos comum foi um vídeo de comédia.

Para uma análise mais específica e a um nível mais baixo, fizemos algumas análises a canais individuais, assim como a vídeos individuais e a datas específicas.

Na Figura 29, fazemos uma análise individual ao canal MrBeast em termos de visualizações à medida do tempo.

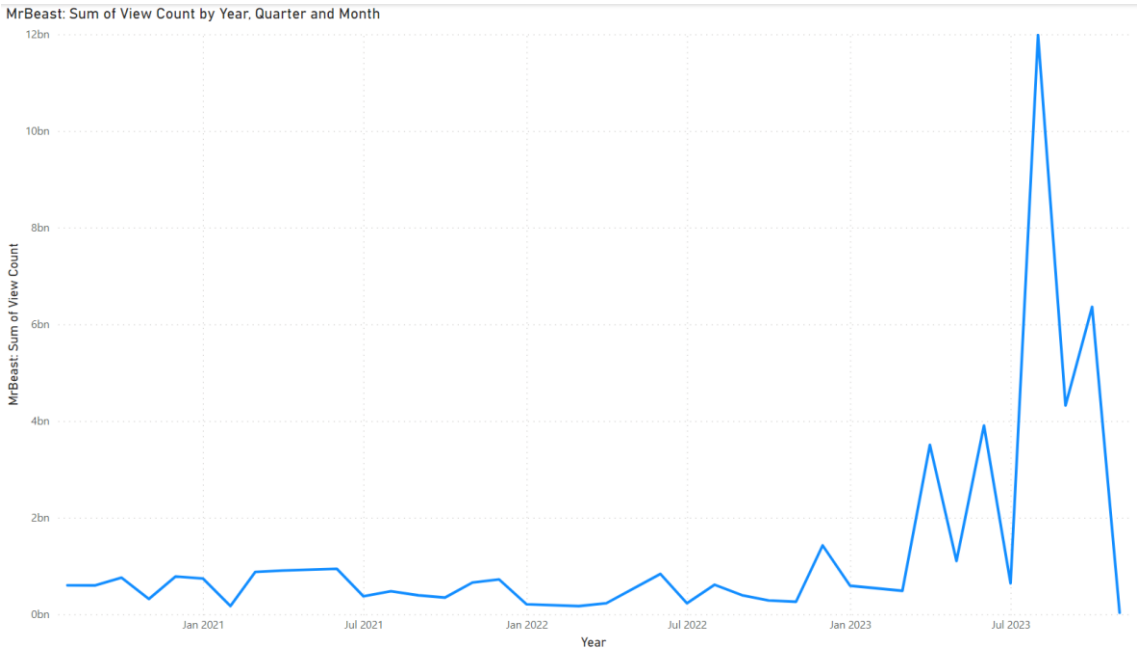

Figura 29 – Número de visualizações para o canal MrBeast à medida do tempo.

Podemos observar que houve uma grande subida em 2023, mais especificamente no Verão. Tal como explicado anteriormente, o Verão é um período de grande atividade no YouTube, especialmente para a categoria do entretenimento.

Na Figura 30 fazemos uma análise semelhante, mas para o canal BLACKPINK, que é um canal de música.

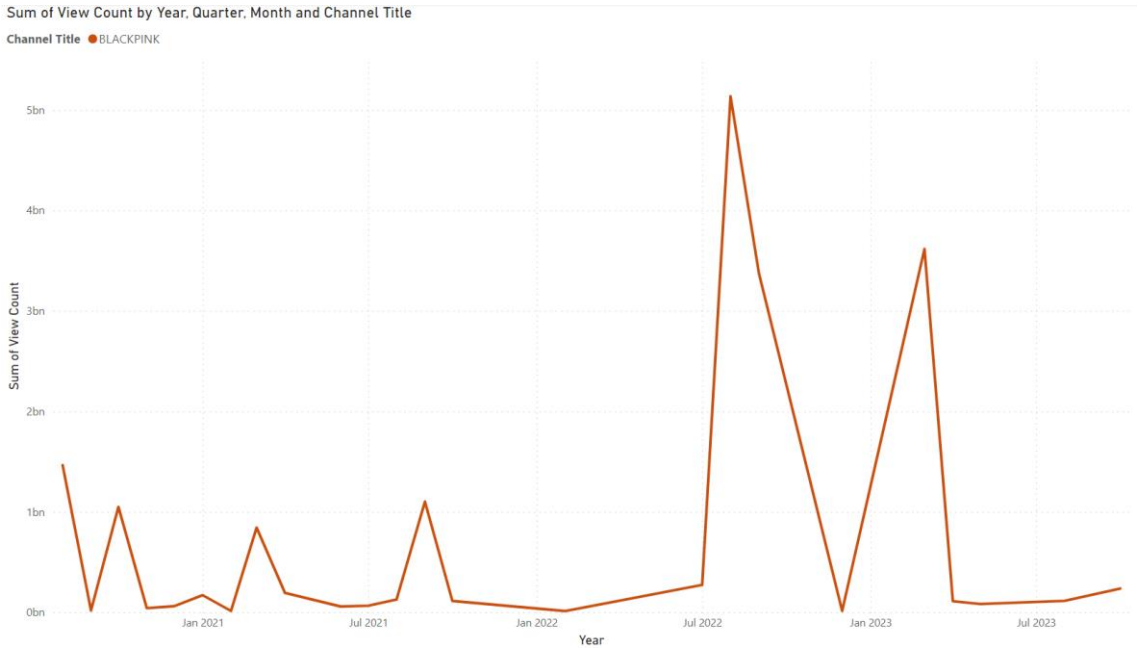

Figura 30 – Número de visualizações para o canal BLACKPINK à medida do tempo.

Como este canal é de música, o período de atividade é mais aleatório, estando relacionado com a qualidade/tendência da música publicada. Portanto, quando um canal está nesta categoria, poderá não ter de se preocupar com a data de publicação, dependendo do contexto.

Após esta análise, fazemos uma comparação entre o ano 2021 e 2022 para o número de visualizações.

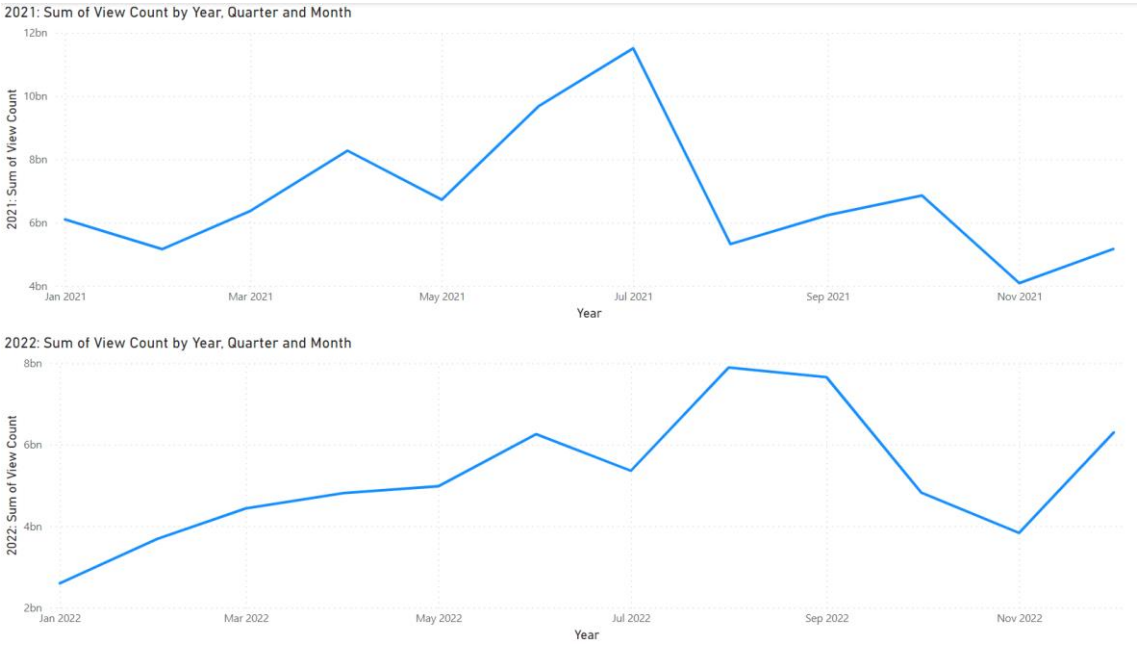

Figura 31 – Comparação entre os anos 2021 e 2022 para o número de visualizações.

Podemos observar uma semelhança entre os dois anos, mas um pequeno atraso no pico do Verão em 2022. No entanto, este pico durou mais tempo.

Existe também uma consistência na queda constante em novembro, e em maio as visualizações crescem sempre, devido ao início da época do verão.

Na Figura 32 fazemos mais uma comparação entre os anos 2021 e 2022, mas desta vez para a hora de publicação.

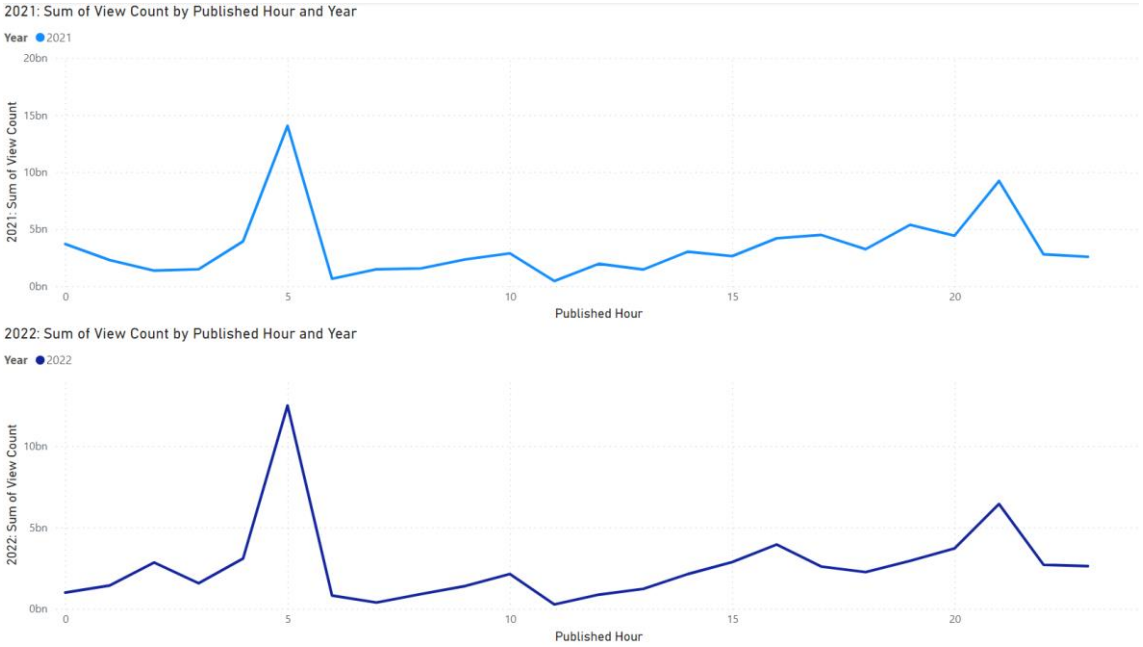

Figura 32 – Comparação entre os anos 2021 e 2022 para a hora de publicação.

Mais uma vez, vemos uma prevalência das 5 horas da manhã na hora de publicação, assim como das 9 horas da noite. No entanto, não vemos uma prevalência como no gráfico geral às 5 horas da tarde, em nenhum destes anos. Assim, podemos concluir que há sempre um número muito elevado de visualizações para vídeos publicados às 5 horas da manhã, sendo esta hora a melhor hora para publicar um vídeo.

Na Figura 33 fazemos uma análise a um vídeo em específico ("*Most Oddly Satisfying Video to watch before sleep*"), que apareceu em #1 nas tendências do dataset.

Verificamos a diferença entre as curvas de gostos e visualizações neste vídeo, que acaba por ter um resultado muito semelhante, apesar do óbvios facto de serem números muito diferentes.

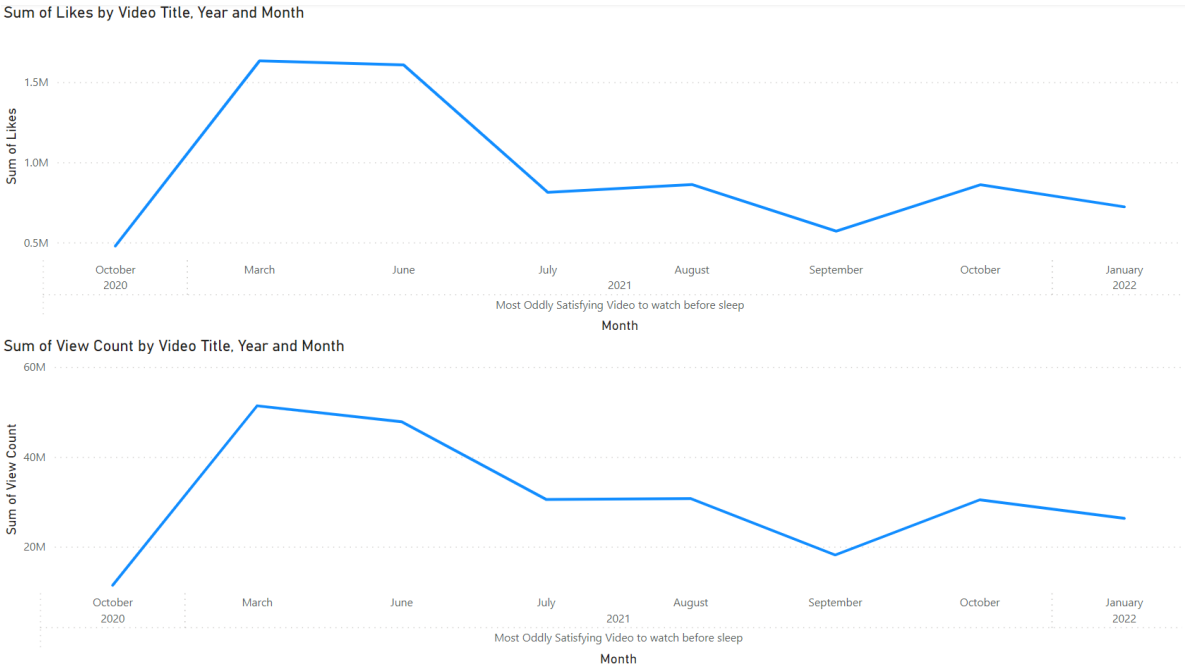

Figura 33 – Comparação entre visualizações e gostos para o vídeo em #1 nas tendências.

Dado o tema do vídeo, é difícil justificar a razão desta tendência. No entanto, notamos um pico entre março e junho quase constante, e uma grande subida assim que o vídeo foi publicado (outubro de 2020).

Para mostrar os gráficos em que havia demasiados dados, como na análise de dados estatísticos de canais ou vídeos, selecionámos os dados visíveis e adicionávamos através da opção “include”, que gera um filtro no PowerBI, conforme na Figura 34.

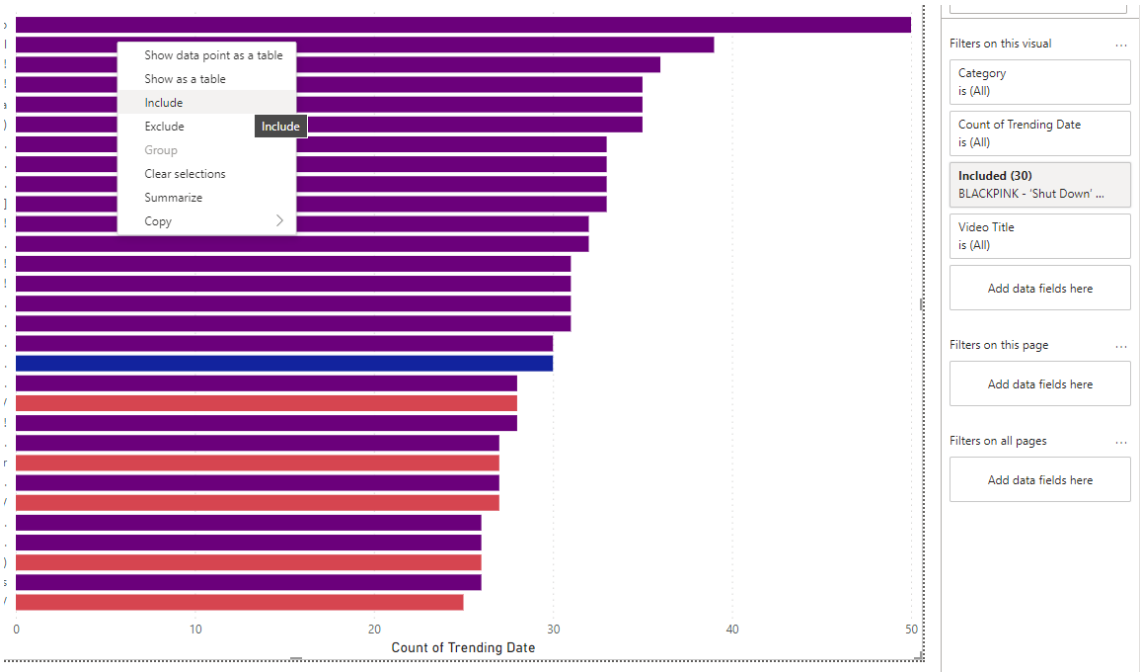

Figura 34 – Opção “include” para filtrar quando há demasiada informação num gráfico.

Uma vez que a informação se encontra ordenada de forma decrescente, bastou-nos selecionar os primeiros valores visíveis.

6. Conclusões e considerações finais

Podemos concluir que este projeto foi muito enriquecedor para os nossos conhecimentos profissionais e académicos, neste caso na área de Sistemas de Informação e Data Analytics.

Tivemos algumas dificuldades na escolha de um tema/dataset, mas estamos satisfeitos com os datasets que encontrámos, por terem muita informação que pode ser analisada de diversas maneiras, como é possível observar na análise OLAP.

Os resultados da análise OLAP foram, a nosso ver, ricos em termos de informação, uma vez que estes permitem-nos de tirar algumas conclusões sobre a plataforma do YouTube e dá-nos alguma informação útil para alguém que queira abrir um canal neste site, como por exemplo, horas de publicação, temas/categorias e algumas tendências da plataforma que podem ajudar na análise futura de dados de qualquer canal.

Na análise OLAP também foram utilizados todos os campos da tabela de facto (FactTrending), mostrando que todos estes campos são necessários para uma análise a esta plataforma, mais especificamente na relação entre visualizações/comentários/gostos/não-gostos e categorias de vídeos, assim como algumas propriedades dos mesmos, como a desativação dos comentários ou dos gostos/não-gostos (*ratings*).

7. Bibliografia

[1] “Most Subscribed YouTube Channels,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data. [Acedido em 7 11 2023].

[2] “YouTube Trending Video Dataset (updated daily),” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset?select=US_youtube_trending_data.csv%29. [Acedido em 23 10 2023].

[3] “YouTube,” Google, [Online]. Available: https://www.youtube.com/. [Acedido em 13 11 2023].

[4] “PowerBI Documentation,” Microsoft, [Online]. Available: https://learn.microsoft.com/en-us/power-bi/. [Acedido em 11 11 2023].

[5] “Python Documentation,” Python, [Online]. Available: https://www.python.org/doc/. [Acedido em 10 11 2023].