Rozbor Youtube trending videí

BIDS Zimní Semestr 2022/23

Erik Prchlík
Tereza Lukschová
Klaudia Balážová

1. ÚVOD A CIEĽ 5

2. METODIKA 6

2.1. VYBRANÉ DATASETY 6
2.2. ZVOLENÉ KPI 7

3. ŠTRUKTÚRA SQL DATABÁZY 8

3.1. DIMENZIONÁLNY MODEL 8
3.2. SQL DATABÁZE 14
3.3. TRANSFORMÁCIA DÁT A NAPLNENIE SQL DATABÁZE 15
3.4. SPRACOVANIE DÁT PRE DIMENZIU KATEGÓRIE 16
3.5. SPRACOVANIE DÁT PRE DIMENZIU A FAKTICKÚ TABUĽKU VIDEA 18
3.6. SPRACOVANIE DÁT PRE DIMENZIU KANÁL 21
3.7. SPRACOVANIE DÁT PRE FAKTICKÚ TABUĽKU KANÁLOV 22
3.8. SPRACOVANIE DÁT PRE DIMENZIU A FAKTOVÚ TABUĽKU ZNAČKA 26

4. PREDPRÍPRAVA VIZUALIZÁCIE DÁT 30

4.1. NAHRANIE A PRÍPRAVA DÁT DO POWER BI 30
4.2. ÚPRAVA DÁT V POWER BI 32

5. VÝSLEDKY 33

5.1. OBĽÚBENOSŤ KATEGÓRIÍ VIDEA V ZÁVISLOSTI OD KRAJINY 33
5.2. POČET TRENDOVÝCH VIDEÍ PRE KANÁL PRE KAŽDÝ MESIAC/ROK 33
5.3. LIKE RATIO TRENDOVÝCH VIDEÍ PRE KANÁL 35
5.4. POMER POČTU TRENDOVÝCH VIDEÍ A CELKOVÉHO POČTU VIDEÍ PRE OBĽÚBENÝ KANÁL 36
5.5. POMER POČTU VIDEÍ A ODBEROV PRE OBĽÚBENÝ KANÁL 37
5.6. POMER POČTU ODBEROV A ZHLIADNUTIE OBĽÚBENÉHO KANÁLU 37
5.7. ZMENY V KATEGÓRIÁCH TRENDING VIDEÍ V ČASE 38
5.8. POČET VIDEÍ A KANÁLOV V KRAJINE 40
5.9. ZASTÚPENIE KANÁLOV V KATEGÓRIÁCH 41

6. DISKUSIA A ZÁVER 42

7. LITERATÚRA 43

8. PRÍLOHY 44

Zoznam Tabuliek

Table 1 - Popis dimenzie video 9
Table 2 - Popis dimenzie značiek 9
Table 3 - Popis dimenzie typu kategórie 10
Table 4 - Popis dimenzie kategória 10
Table 5 - Popis dimenzie kanálu 10
Table 6 - Popis dimenzie krajina 11
Table 7 - Popis dimenzie dátum 11
Table 8 - Popis faktickej tabuľky značiek 12
Table 9 - Popis faktickej tabuľky videa 12
Table 10 - Popis faktickej tabuľky kanálu 13

Zoznam Obrázkov

Obrázok 1: Dimenzionálny model Youtube problematiky. 8
Obrázok 2: SQL skript pre vytvorenie dimenzie videa. 14
Obrázok 3: SQL skript pre vytvorenie faktickej tabuľky pre značku videa 15
Obrázok 4: SQL skript pre naplnenie dimenzie dátumami v požadovanom formáte. 15
Obrázok 5: Podoba dát v priebehu spracovania dát kategórií videí. 16
Obrázok 6: Skript v programovacom jazyku Python pre spracovanie kategórií do požadovanej formy. 17
Obrázok 7: SSIS nahrávajúci dáta do dimenzie kategória. 17
Obrázok 8: CSV súbor datasetu trending videí otvorený v progrem Microsoft Excel. 18
Obrázok 9: Nastavenie súborového manažéra v programe Microsoft SQL Server Data Tools, pre čítanie CSV súboru trending videí. 19
Obrázok 10: Výraz použitý v Derived column pre úpravu dátumov do požadovanej formy. 19
Obrázok 11: SSIS nahrávajúci dáta do dimenzie video. 20
Obrázok 12: SSIS nahrávajúci dáta do faktickej tabuľky video. 20
Obrázok 13: SQL skript riešiaci problém duplicity vo faktickej tabuľke video. 21
Obrázok 14: CSV súbor datasetu Top 1000 youtuberov otvorený v progrem Microsoft Excel.ä 21
Obrázok 15: Popisné dáta pre dimenziu video, nachádzajúce sa v datasete trending videí. 22
Obrázok 16: SSIS nahrávajúci dáta do dimenzie kanál. 22
Obrázok 17: Ukážka nevhodných dát, ktoré bolo nutné ošetriť. 23
Obrázok 18: Prvý krok pri otváraní JSON súboru v programe Microsoft Excel. 23
Obrázok 19: Po prečítaní JSON súboru, je získaný list záznamov prekonvertovaný na tabuľku. 24
Obrázok 20: Získaná tabuľka s komplexným stĺpcom obsahujúcim všetky atribúty záznamu. 24
Obrázok 21: Rozdelenie komplexného stĺpca na požadovaná atribúty ako stĺpce. 24
Obrázok 22: Finálna tabuľka pripravená pre potrebnú transformáciu. 25
Obrázok 23: Komplexný SSIS načítavajúci dáta z dvoch zdrojov pre faktickú tabuľku kanál. 25
Obrázok 24: Postup pri získavaní identifikačného čísla krajiny a ošetrenie záznamov s nenájdenými ID. 26
Obrázok 25: Vylúčenie kanálov z JSON súboru pri nenájdení zhody a ošetrenie chýbajúcich atribútov. 26
Obrázok 26: Ošetrenie chýbajúcich a neaktuálnych údajov odberateľov kanálu.ä 26
Obrázok 27: Vyextrahované značky videí z datasetu trending videí. 27
Obrázok 28: Skript v programovacom jazyku Python pre spracovanie značiek videí. 27

Obrázok 29: SQL skript pre vymazanie duplicitných záznamov v bezfakticko faktickej tabuľke značiek. 28
Obrázok 30: SSIS nahrávajúci dáta do dimenzie značka. 28
Obrázok 31: SSIS nahrávajúci dáta do bezfaktickej faktickej tabuľky značka. 29
Obrázok 32: Script pre vytvorenie ViewVideo 30
Obrázok 33: Script pre vytvorenie ViewKanál 31
Obrázok 34: Model v PowerBI 32
Obrázok 35: Počet videí podľa kategórie videa v krajine 33
Obrázok 36: Zapätie pre 2 a 3 KPI 34
Obrázok 37: Počet trendových videí pre všetky kanále za každý mesiac a rok. 34
Obrázok 38: Počet trendových videí pre všetky kanále za každý mesiac. 34
Obrázok 39: Počet videí počas a po koronavírusovom období 35
Obrázok 40: like ratio v čase pre kanál Vijay Television 35
Obrázok 41: Like ratio pre kategóriu v čase 36
Obrázok 42: vizualizácia pomera počtu videí a celkového počtu videí kanálu 37
Obrázok 43: vizualizácia popularity 38
Obrázok 44: vizualizácia zastúpenia kategórií v krajinách počas koronavírusových opatrení 39
Obrázok 45: vizualizácia zastúpenia kategórií v krajinách po koronavírusových opatreniach 39
Obrázok 46: zmena počtu trenodvých videí označených kategóriou „správy“ v čase 40
Obrázok 47: vizuál pre zobrazenie zastúpenia počtu trendových vidiei v mape 41
Obrázok 48: vizuál pre zobrazenie zastúpenia kanálov podľa kategórie videa 41

1. Úvod a cieľ

Youtube.com je v momentálnej chvíli najväčšia sociálna platforma pre nahrávanie, zdieľanie a prehrávanie videí. Podľa oficiálneho YouTube blogu (2017), k roku 2017 trávili používatelia dohromady sledovaním videí až miliardu hodín denne. Po koronavírusovej kríze, keď muselo ľudstvo tráviť viac času doma a odvetvie internetovej zábavy iba prekvitalo, veríme, že k dnešnému dňu by toto číslo mohlo, byť násobne väčšie.

Platforma sa počas rokov stala aj miestom, kde si každý tvorca obsahu môže zarobiť peniaze. Na základe živnosti, zmluvy so samotným YouTube a nastavení reklám sa každý môže stať tzv. "youtuberom" - úspešnosť tejto kariéry však veľmi stojí na publiku YouTube. Ide o počet a dobu zhliadnutia, angažovanosť publika alebo aj počet fanúšikov, ktorý videá kanálu sledujú pravidelne. Pre samotné získanie zmluvy s YouTube je nutné splniť niekoľko podmienok – k roku 2023 sú podmienky napríklad získanie 1000 odberateľov (alebo fanúšikov, ktorí sa zaujímajú o tvorbu umelcov) alebo dokonca 1 milión zhliadnutí videí Shorts v posledných 90tich dňoch (Youtube Help, 2023).

Pre budúcich aj súčasných tvorcov obsahu je pre úspech veľmi dôležité zapojenie publika. To sa však odvíja od samotného obsahu. Videá sa užívateľom ponúkajú podľa zložitého algoritmu, ktorý zahŕňa nielen záujmy používateľov (alebo typy videí, na ktoré bežne pozerá a ktoré by ho mohli zaujímať či baviť) tak aj samotnú popularitu videí. Momentálne populárne videá pre územie, v ktorom sa divák nachádza, sa ponúka v záložke "Trendy" na hlavnej stránke. Umiestnenie medzi týmito videami zaistia tvorcom veľké rozšírenie publika, a tak viac možností a menej obmedzení čo sa týka budúcej tvorby.

Náš projekt sa bude zameriavať práve na tieto trendy videa a samotných tvorcov na YouTube. Budú skúmané videá, ktoré sa dostali do zoznamu trendových a obľúbené kanály platformy. Cieľom projektu je zobrazenie zaujímavostí spojených s týmito dvoma kategóriami z dvoch rôznych pohľadov.

Prvým z pohľadov sú zaujímavosti, ktoré môžu byť dôležitým ukazovateľom pre nových alebo aj súčasných tvorcov video obsahu na tejto platforme. Venovať sa budeme otázkam ako napríklad: „aké ľudia radi pozerajú videá?“, „v ktorej krajine aké kategórie videí ľudia pozerajú najčastejšie?“, „ktoré videá sa dostávajú do trendov?“, „koľko z trendových videí je naozaj dobrých, alebo ich ľudia pozerajú len pretože sú bizarné?“ a mnoho ďalších otázok, z ktorých výsledky im môžu pomôžu sa inšpirovať alebo zlepšovať v tomto prekvitajúcom odvetví.

Ďalším pohľadom sú zaujímavosti týkajúce sa týchto videí a kanálov v období počas (do konca mája 2021) a po koronavírusových opatrení (od prvého júna 2021). Venovať sa budeme otázkam ako sú napríklad: „aké ľudia pozerali videá, a aké pozerajú teraz?“, „zaujímali sa počas koronavírusových opatrení viac o správy, ako dnes?“, „koľko videí priemerne vydávali kanále a koľko teraz?“ alebo „vychádzali videá s lepším pomerom likov k dislikom ako vychádzajú dnes?“.

Taktiež sa budeme v práci snažiť, aj vďaka týmto definovaným otázkam, potvrdiť alebo nepotvoriť naše predom stanovené hypotézy.

2. Metodika

2.1. Vybrané datasety

Základným medzníkom úspechu v tomto projekte bolo správne vybratie dát, s ktorými sa bude pracovať. Pre nás základným zdrojom dát bola webová stránka Kaggle, ktorá je online komunitou zaoberajúcou sa dátami a strojovým učením (Kaggle Youtube Kanál, 2019). Táto platforma ponúkla široké rozpätie datasetov týkajúcich sa Youtube užívateľov, Youtube videí a trending videí na Youtube. Mnohé z nich opakovali rovnaké dáta, ich kombináciou bolo však v rámci možností možné získať kvalitný prehľad.

YouTube Trending Video Dataset (updated daily)
Zdroj dat: https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset?select=GB_youtube_trending_data.csv

Dataset "YouTube Trending Video Dataset (updated daily)" v preklade "Dataset Youtube obľúbených videí (denne aktualizované)" zodpovedá svojmu názvu. Ide o dataset, ktorý každý deň zaznamenáva "trending videa" v ten istý konkrétny deň. Jedno to isté video sa teda môže v datasete vyskytovať viac krát – či už v neprerušenom intervale (video je "trending" neustále v spojitý čas) alebo naopak v intervale prerušenom (video bolo "trending" a po čase sa do "trending" znovu dostalo). Začiatok datasetu sa datuje ku dňu 12.8. 2020. Ku dňu 26.10.2022 dataset obsahoval celkom 29 720 záznamov.

Dataset obsahuje CSV súbory s dátami o videách a JSON súbory popisujúce definované kategórie videí – oba súbory vždy pre konkrétnu krajinu. Krajinami v datasete sú Brazília, Kanada, Nemecko, Francúzsko, Veľká Británia, India, Japonsko, Kórea, Mexico, Rusko a Spojené Štáty Americké. Rozlišovanými atribútmi sú napríklad názvy kanálov, názvy videí, trending dátum, kategória videa, počet palcov nahor, počet palcov nadol, počet komentárov alebo napríklad zoznam značiek (tzv. "hashtagov"), pod ktorými možno video nájsť.

YouTube's Channels Dataset
Zdroj dat: https://www.kaggle.com/datasets/harshithgupta/youtubes-channels-dataset

Dataset "YouTube's Channels Dataset" v preklade "Dataset Youtube kanálov" obsahuje dáta ohľadom tzv. kanálov na Youtube. Kanálom sa označuje prihlásený používateľ youtube, ktorý tvorí videá. Ide o analytický dataset, v ktorom už autor porovnáva voči sebe základné atribúty kanálu. Cieľom použitia týchto dát je získanie dodatočných informácií o kanáloch a jednoduchých štatistík, ktoré by sa prepojili s "trending videí" z predchádzajúceho datasetu.

Ku dňu 26.10.2022 obsahuje dataset okolo 11500 záznamov s presne 26 atribútmi pre každý zo záznamov. Príklady atribútov sú napríklad pomer celkového zhliadnutia všetkých videí voči vypnutiu akéhokoľvek videa z kanála, počet prehliadnutí kanála, počet videí, pomer palcov nahor či palcov nadol voči počtu odberateľov alebo napríklad súčet palcov nahor, komentárov a ďalších.

Top 1000 Youtubers (Cleaned) World
Zdroj dat: https://www.kaggle.com/datasets/syedjaferk/top-1000-youtubers-cleaned

Dataset „Top 1000 Youtubers (Cleaned) World“ vo voľnom preklade „1000 najlepších Youtuberov sveta“ je ďalší dataset obsahujúci informácie o kanáloch na platforme Youtube. Dáta sú zoradené od najúspešnejšieho kanála (podľa atribútu “Rank”) a ukazujú tak užitočné informácie. Posledná aktualizácia datasetu prebehla v júni 2022 a ako už názov napovedá nachádza sa v ňom 1000 záznamov – jeden vždy pre každý kanál.

Atribúty dát v tomto sete sú nielen rank, názov kanálu alebo kategórie, ktorej sa kanál venuje. Získané dáta sú o niečo všeobecnejšie ako v predchádzajúcom datasete, ale poskytujú nám nové informácie ako napríklad cieľovú krajinu publika.

2.2. Zvolené KPI

KPI alebo „kľúčové ukazovatele výkonnosti“ sú skratka anglického originálu „key performance indicators“ (Hankusová, 2020). Ide o sprostredkovateľa na meranie výkonnosti, ktorý sa bežne používa na získanie prehľadu o úspešnosti konkrétnej aktivity organizácie. Tím Microsoftu 365 (2019) označuje KPI ako spoľahlivé indikátory zdravia spoločnosti. Vďaka KPI môže firma sledovať úspech a zlyhanie svojich obchodných taktík, aby tak mohla trvalo zlepšovať a stavať na myšlienkach, ktoré firmu posúvajú dopredu k úspechu.

V našom prípade sú dôležité nie len pri sledovaní výsledkov z nami zvolených otázok a skutočnosti, či sa výsledky približujú k predpokladu, ale aj pri inšpirácií a pomoci pre nových alebo súčasných tvorcov na platforme Youtube.

Pre náš projekt sme vyberali nasledujúce KPI:

1. Obľúbenosť kategórií videa v závislosti na krajine.
   Hypotéza: Má kategória zábavných videí viac ako 30% zastúpenie v každej krajine?
2. Počet trendových videí pre kanál pre každý mesiac/rok.
   Hypotéza: Počet videí na kanál sa počas koronavírusu zvýšil o viac ako 10 %.
3. Like ratio trendových videí pre kanál.
4. Pomer počtu trendových videí a celkového počtu videí pre obľúbený kanál.
5. Pomer počtu videí a odberov pre obľúbený kanál.
6. Pomer počtu odberov a zhliadnutí obľúbeného kanálu.
7. Zmeny v kategóriách trending videí v čase.
   Hypotéza: Zvýšenie počtu trending videí kategórie správ v období koronavírusu.

3. Štruktúra SQL databázy

3.1. Dimenzionálny model

Pre uchovanie získaných dát bol vytvorený dimenzionálny model. Kvôli priestorovým možnostiam nie je v dokonalom tvare vločky, ale opticky je možné vidieť jej náznaky. Tento model sa skladá z dvoch faktických tabuliek, jednej bezfaktickej faktickej tabuľky a siedmich dimenzionálnych tabuliek. Bezfaktická faktická tabuľka predstavuje takzvaný mostový spoj medzi dimenziou video a značka. Dimenzionálne tabuľky prevažne obsahujú popisné informácie objektu, s výnimkou časovej dimenzie. Na druhú stranu faktické uchovávajú z pravidla číselné údaje.
Dimenzionálny model vo väčšom meradle je tiež súčasťou našej prílohy.

Obrázok 1: Dimenzionálny model Youtube problematiky.

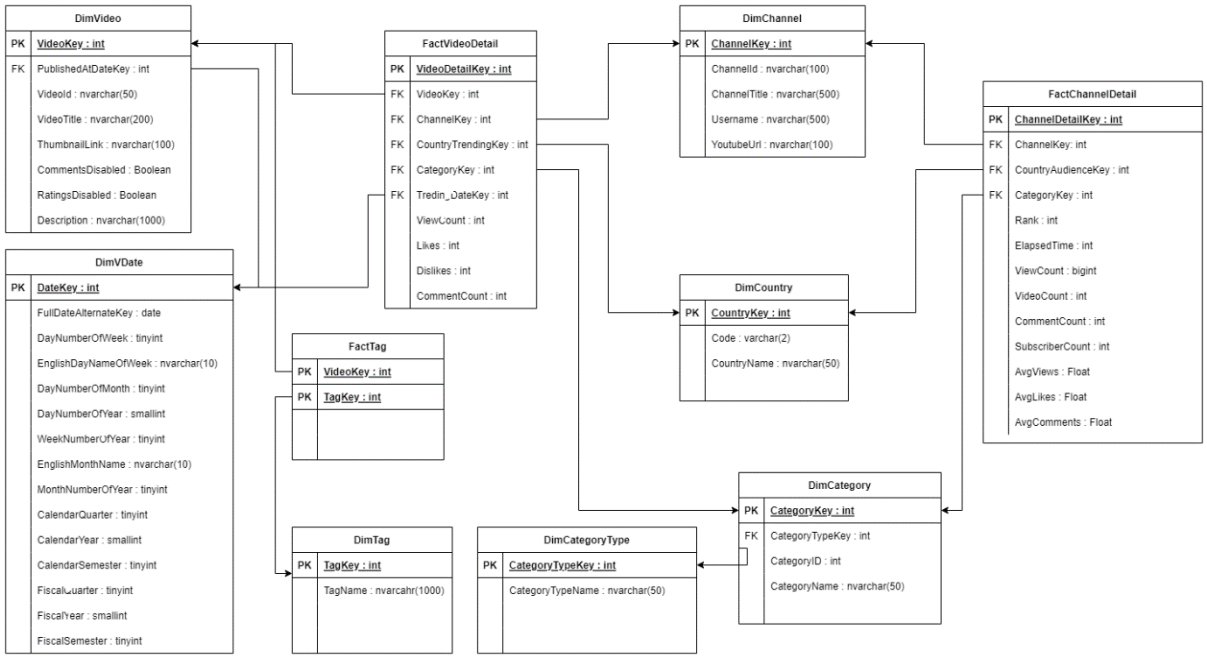


Table 1 - Popis dimenzie video

DimVideo							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	VideoKey	int		identity	VideoKey. Primary key for table DimVideo		
2.FK	PublishedAtDateKey	int			Key for video published date. ('	1	')
3.	VideoId	nvarchar(50)			The Id of the video. ('	J78aPJ3VyNs	')
4.	VideoTitle	nvarchar(200)	ok		Title of video. ('	I left youtube for a month and THIS is what happened.	')
5.	ThumnailLink	nvarchar(100)	ok		Youtube link. ('	https://i.ytimg.com/vi/J78aPJ3VyNs/default.jpg	')
6.	CommentsDisabled	Boolean	ok		Video have disabled comments. ('	FALSE	')
7.	RatingsDisabled	Boolean	ok		Video have disabled ratings. ('	FALSE	')
8.	Description	nvarchar(1000)	ok		Description of video. ('	I left youtube for a month and this is what happenedMY	')

Table 2 - Popis dimenzie značiek

DimTag							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	TagKey	int		identity	TagKey. Primary key for table DimTag		
2.	TagName	nvarchar(1000)	ok		Name of tag. ('	funny meme	')

Table 3 - Popis dimenzie typu kategórie

DimCategoryType							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	CategoryTypeKey	int		identity	CategoryTypeKey. Primary key for table DimCategoryType		
2.	CategoryTypeName	nvarchar(50)	ok		Category type of dimension. ('	Video	')

Table 4 - Popis dimenzie kategória

DimCategory							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	CategoryKey	int		identity	CategoryKey. Primary key for table DimCategory		
2.FK	CategoryTypeKey	int			Key to CategoryType. ('	1	')
3.	CategoryId	int			Id of category from source dataset. ('	1	')
4.	CategoryName	nvarcgar(50)	ok		Category name of video or channel. ('	Channel	')

Table 5 - Popis dimenzie kanálu

DimChannel							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	ChannelKey	int		identity	ChanelKey. Primary key for table DimChannel		
2.	ChannelId	nvarcahr(100)			The Id of the channel. ('	UCdzU3DSGzyWzN2118yd9X9g	')
3.	ChannelTitle	nvarchar(500)	ok		Title of channel. ('	jacksepticeye.	')
5.	Username	nvarchar(500)	ok		Username/handle of the account. ('	tseries	')
6.	YoutuberUrl	nvarchar(100)	ok		URL like for the account ('	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy4_brA	')

Table 6 - Popis dimenzie krajina

DimCountry							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	CountryKey	int		identity	CountryKey. Primary key for table DimCountry		
2.	Code	varchar(2)	ok		Code of country. ('	GB	')
3.	CountryName	nvarchar(15)	ok		Whole name of country. ('	Great Britain	')

Table 7 - Popis dimenzie dátum

DimDate							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	DateKey	int		identity	DateKey. Primary key for table DimDate		
2.	FullDateAlternateKey	Date			Full date in formate FFFF-MM-DD. ('	2022-12-31	')
3.	DayNumberOfWeek	tinyint	ok		Day number of week. ('	1	')
4.	EnglishDayNameOfWeek	nvarchar(10)	ok		English day name of week. ('	Monday	')
5.	DayNumberOfMonth	tinyint	ok		Day number of month. ('	1	')
6.	DayNumberOfYear	smallint	ok		Day number of year. ('	1	')
7.	WeekNumberOfYear	tinyint	ok		Week number of year. ('	1	')
8.	EnglishMonthName	nvarchar(10)	ok		English month name. ('	January	')
9.	MonthNumberOfYear	tinyint	ok		Month number of year. ('	1	')
10.	CalendarQuarter	tinyint	ok		Calendar quarter. ('	1	')
11.	CalendarYear	smallint	ok		Calendar year. ('	2022	')
12.	CalendarSemester	tinyint	ok		Calendar semester. ('	1	')
13.	FiscalQuarter	tinyint	ok		Fiscal quarter. ('	1	')
14.	FiscalYear	smallint	ok		Fiscal year. ('	2022	')
15.	FiscalSemester	tinyint	ok		Fiscal semester. ('	1	')

Table 8 - Popis faktickej tabuľky značiek

FactTag					
Key	Name	Data Type	Null	Attributes	Description
1.PK	VideoKey	int		identity	Key to connect tag with video.
2.PK	TagKey	int		identity	Key to connect video with tag.

Table 9 - Popis faktickej tabuľky videa

FactVideoDetail							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	VideoDetailKey	int		identity	VideoDetailKey. Primary key for table FactVideoDetail		
2.FK	VideoKey	int			The key of the video. ('	J78aPJ3VyNs	')
3.FK	ChannelKey	int			The key of the channel. ('	1	')
4.FK	CountryTrendingKey	int			Key of country where video was in trendings. ('	1	')
5.FK	CategoryKey	int			The key of the category. ('	1	')
7.FK	TendingDateKey	int			Key for date in which video was trending. ('	1	')
8.	ViewCount	int	ok		Number of video views. ('	100	')
9.	Likes	int	ok		Number of likes. ('	100	')
10.	Dislikes	int	ok		Number of dislikes. ('	578	')
11.	CommentCount	int	ok		Number of comments. ('	23456	')

Table 10 - Popis faktickej tabuľky kanálu

FactChannelDetail							
Key	Name	Data Type	Null	Attributes	Description		
1.PK	ChannelDetailKey	int		identity	VideoDetailKey. Primary key for table FactChannelDetail		
2.FK	ChannelKey	int			The key of the channel. ('	1	')
3.FK	CountryAudienceKey	int			Key of channel main audience country. ('	1	')
4.FK	CategoryKey	int			The key of the category. ('	1	')
5.	Rank	int	ok		Rank of the channel. ('	1	')
6.	ElapsedTime	int	ok		Chanel elapsed time. ('	356654	')
7.	ViewCount	int	ok		Number of likes. ('	100	')
8.	VideoCount	int	ok		Number of dislikes. ('	578	')
9.	CommentCount	int	ok		Number of comments. ('	23456	')
10.	SubscriberCount	int	ok		Total number of subscribers. ('	220100000	')
11.	AvgViews	float	ok		Average views of the content. ('	57000	')
12.	AvgLikes	float	ok		Average likes of the videos/content. ('	1700	')
13.	AvgComments	float	ok		Average Comments of the video/content. ('	108	')

3.2. SQL databáze

Za pomoci Microsoft SQL Server Management Studio, s využitím SQL serveru Treeman, boli vytvorené tabuľky SQL databázy na základe dimenzionálneho modelu zobrazeného vyššie. Dátové typy atribútov jednotlivých tabuliek boli priebežne upravované v priebehu procesu transformácie dát z dôvodu problematických situácií pri ukladaní dlhých reťazcoch v kódovaní UTF-8.

Obrázok 2: SQL skript pre vytvorenie dimenzie videa.

```sql
USE [AW_DWH_xprchlik]
GO

/****** Object:  Table [dbo].[dimVideo]    Script Date: 13.01.2023 16:19:44 ******/
SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE [dbo].[dimVideo](
    [VideoKey] [int] IDENTITY(1,1) NOT NULL,
    [VideoId] [dbo].[Name] NULL,
    [VideoTitle] [nvarchar](200) NULL,
    [PublishedAtDateKey] [int] NOT NULL,
    [ThumbnailLink] [nvarchar](100) NULL,
    [CommentsDisabled] [bit] NULL,
    [RatingDisabled] [bit] NULL,
    [Description] [nvarchar](1000) NULL,
 CONSTRAINT [PK_dimVideo] PRIMARY KEY CLUSTERED
(
    [VideoKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE =
OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

ALTER TABLE [dbo].[dimVideo]  WITH CHECK ADD  CONSTRAINT [FK_dimVideo_dimVDate]
FOREIGN KEY([PublishedAtDateKey])
REFERENCES [dbo].[dimVDate] ([DateKey])
GO

ALTER TABLE [dbo].[dimVideo] CHECK CONSTRAINT [FK_dimVideo_dimVDate]
GO
```

Priložený skript pre vytvorenie dimenzionálnej tabuľky videa je príkladom toho ako boli vytvárané zvyšné tabuľky z navrhovaného modelu. Uložené boli na účet používateľa xprchlik. Samozrejmosťou sú zhodujúce sa atribúty s príslušnými dátovými typmi a pridelenými primárnymi a cudzími kľúčmi.

Pri vytváraní dimenzionálnych tabuliek s minimom záznamov, ako boli dimenzie typ kategórie a krajiny, boli aj zároveň manuálne zaplnené na základe zvolených dátasetov.

Konkrétne krajiny podľa názvov Špecifikom bola bezfaktová faktová tabuľka prepájajúca videá a ich značky.

Obrázok 3: SQL skript pre vytvorenie faktickej tabuľky pre značku videa

```sql
USE [AW_DWH_xprchlik]
GO

/****** Object:  Table [dbo].[factTag]    Script Date: 13.01.2023 12:14:14 ******/
SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE [dbo].[factTag](
    [VideoKey] [int] NOT NULL,
    [TagKey] [int] NOT NULL,
 CONSTRAINT [pk_tagKey_videoKey] PRIMARY KEY CLUSTERED
(
    [TagKey] ASC,
    [VideoKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE =
OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
```

Okrem skriptov pre vytváranie tabuliek je tiež zaujímavé spomenúť ako boli pridané nové dátumy do tabuľky časovej dimenzie nižšie.

Obrázok 4: SQL skript pre naplnenie dimenzie dátumami v požadovanom formáte.

```sql
CREATE TABLE #Dates(Dates DATE)

DECLARE @Start DATE
DECLARE @End DATE

SET @Start ='01/01/2005' -- IN MM/DD/YYYY format
SET @End ='12/31/2022' -- IN MM/DD/YYYY format

DECLARE @CounterINT, @TotalCountINT
SET @Counter = 0
SET @TotalCount = DateDiff(DD,@Start,@End)

WHILE (@Counter <=@TotalCount)
BEGIN
  DECLARE @DateValue DATE
  SET  @DateValue= DATEADD(DD,@Counter,@Start)

    INSERT INTO dimVDate (FullDateAlternateKey)
    VALUES(@DateValue)

    SET @Counter = @Counter + 1
END
```

3.3. Transformácia dát a naplnenie SQL databáze

Každý z projektov, ktorý sa zaoberá analýzou veľkých dát, na začiatku zápasí s úpravou dát do použiteľnej formy. Kvalita dát je kľúčový faktor, ak je nedostatočná, stojí to množstvo času a práce pre dosiahnutie prijateľnej formy. Avšak, ak je problém s kvantitou dát počas spracovania projektu, to je ešte nepríjemnejšie. V tomto projekte nastali oba situácie, ktoré boli riešené do maximálnej možnej miery.

Použité nástroje:

1. Programovací jazyk Python,
2. Microsoft Excel,
3. Sublime Text,
4. Microsoft SQL Server Data Tools.

3.4. Spracovanie dát pre dimenziu kategórie

https://colab.research.google.com/drive/1RMFLDk0UeCk9-gDOWMmxyl0Gw-cn3Odw?usp=sharing

Súčasťou datasetu s trend videami boli aj JSON súbory obsahujúce zoznamy kategórií pre každú sledovanú krajinu zvlášť. Všetky súbory boli prečítané za pomoci programovacieho jazyku Python v prostredí Google Colaboratory a uložené v požadovanej v štruktúre do slovníku.

Obrázok 5: Podoba dát v priebehu spracovania dát kategórií videí.

```
[11] categiryData = readCategoryData(files)

     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
     {'1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '15': 'Pets & Animals', '17': 'Spo
```

Následne boli overené konflikty identifikačných čísiel a prípadné duplicity. Na záver bol unikátny zoznam zapísaný do CSV súboru.

Obrázok 6: Skript v programovacom jazyku Python pre spracovanie kategórií do požadovanej formy.

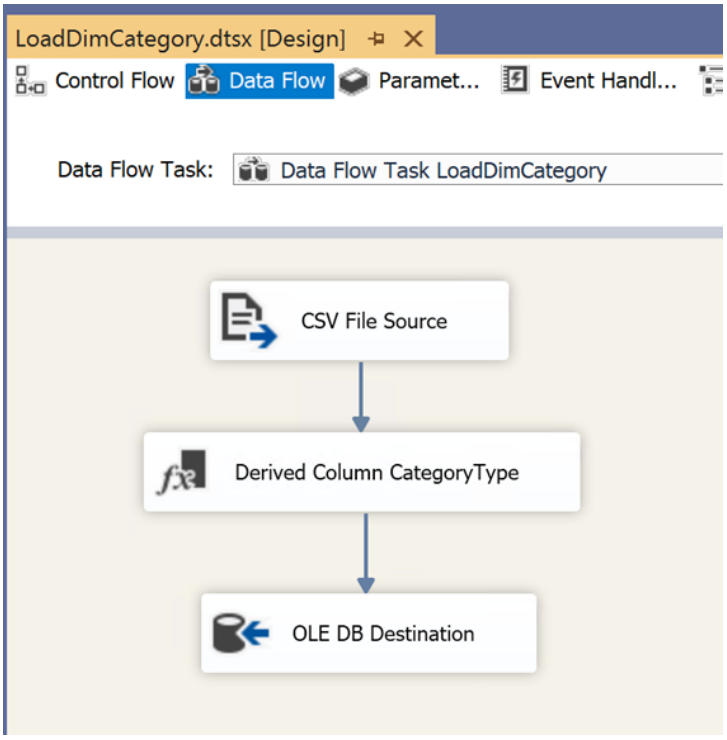
```
[12] import keyword
     def asGBData(data: dict):
         dataGB = data['GB']['items'].copy()
         for country, categories in data.items():
             if dataGB == categories['items']:
                 data[country]['asGB'] = True
             else:
                 data[country]['asGB'] = False
```

```
asGBData(categiryData)
listOfBool = [categories['asGB'] for country, categories in categiryData.items()]
finalCategoryDict = categiryData['GB']['items']
if all(listOfBool) == False:
  for country, categories in categiryData.items():
    if categories['asGB'] == False:
      value = { k : categories['items'][k] for k in set(categories['items']) - set(categiryData['GB']['items']) }
      print(value)
      finalCategoryDict.update(value)
```

Informácie o kategóriách kanálov boli iba v sobore o Top 1000 YouTuberoch za určité obdobie. Pre účely nahratia do databázy bol tento stĺpec vyňatý a očistený od duplicít za pomoci programu Excel a zároveň boli pridané aj ich identifikačné čísla, kvôli ich absencii.

Po úprave dát bolo možné jednoduché uloženie dát do databáze za pomoci Microsoft SQL Server Data Tools. Jedinou potrebnou informáciou, ktorá bola pridaná za pomoci Derived column, bolo identifikačné číslo typu kategórie. Tieto typy boli vložené do databáze manuálne, kde sa rozlišuje, či ide o kategóriu pre vide alebo kanál. Oba CSV súbory s kategóriami boli nahrávané do databázy samostatne.

Obrázok 7: SSIS nahrávajúci dáta do dimenzie kategória.



Celkovo bolo pridaných 116 záznamov spolu s neznámou kategóriou.

3.5. Spracovanie dát pre dimenziu a faktickú tabuľku videa

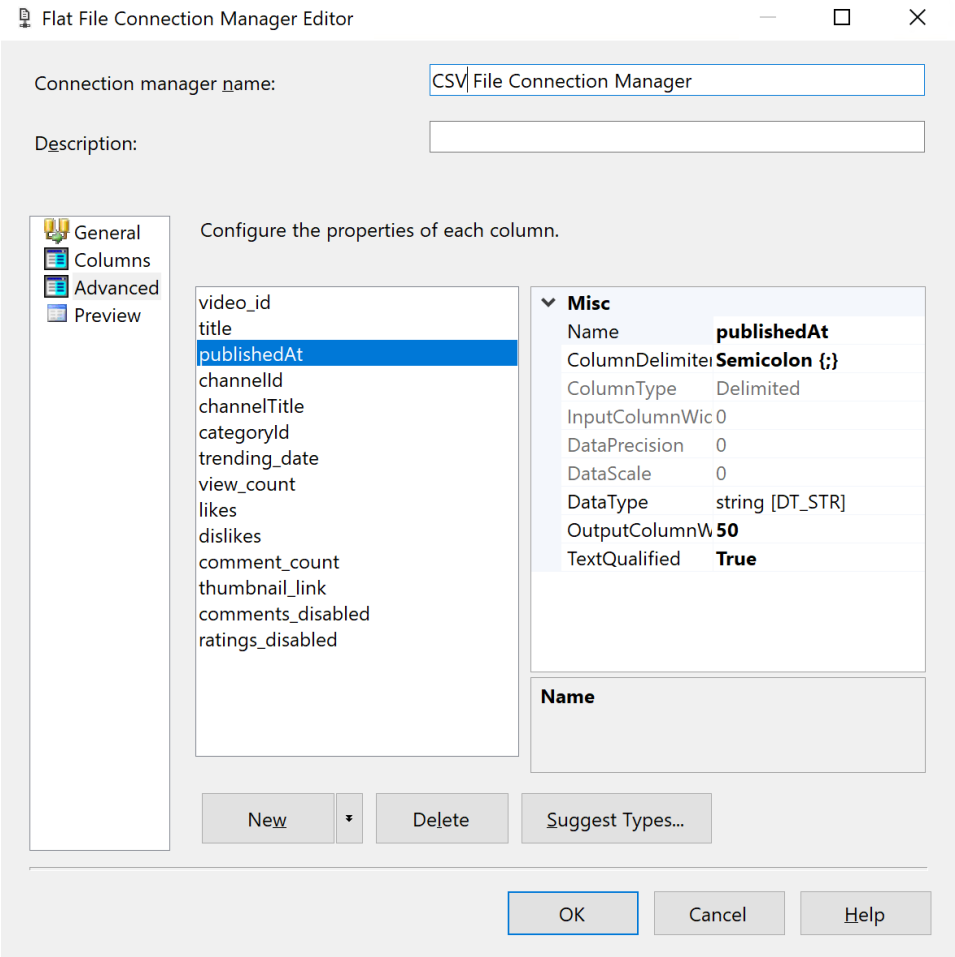
Dáta pre dimenziu Videá pochádzajú zo skupiny súborov o trendových videách. Kvôli nevhodným znakom boli vyradené súbory krajiny Japonsko, Kórea a Rusko. Základným krokom pri tomto spracovaní bolo nahradenie všetkých výskytov bodkočiarky za prázdny znak a zaistenie tohto znaku ako oddeľovač stĺpcov pre tieto súbory. Bolo to kvôli tomu, že oddeľovače pre názvy stĺpcov a medzi záznamami v riadkoch boli rozdielne a program pre nahrávanie do databázy tieto súbory nevedel rozložiť. Takže zložitosť práce s týmito súbormi už od začiatku viedla k práci v programe Excel. Následne bol vyňatý stĺpec so značkami videí, ktoré sa spracovávali samostatne. Súbory sa spojili dokopy a odstráli sa duplicity pre dimenzionálnu tabuľku. Pre faktovú tabuľku sa zase pridal do iného súhrnného súboru stĺpec s kódom krajiny, podľa toho z ktorého súboru príslušný záznam pochádzal. Jediným problémom pri práci v Exceli, bolo to množstvo záznamov pre faktickú tabuľku. Jedná sa o skoro 1,2 milióna záznamov.

Obrázok 8: CSV súbor datasetu trending videí otvorený v progrem Microsoft Excel.

	A	B	C	D	E	F	G
1	video_id	title	publishedAt	channelId	channelTitl	categoryId	trending_
2	s9FH4rDMvds	LEVEI UM FORA?	12.8.2020 0:21	UCGfBwrCc	Pietro Guec	22	12.8.202
3	jbGRowa5tIk	ITZY “Not Shy” M	11.8.2020 17:00	UCaO6TYtlı	JYP Enterta	10	12.8.202
4	3EfkCrXKZNs	Oh Juliana PARÓ	10.8.2020 16:59	UCoXZmVn	As Irm?s M	22	12.8.202
5	gBjox7vn3-g	Contos de Runet	11.8.2020 17:00	UC6Xqz2pr	League of L	20	12.8.202
6	npoUGx7UW7o	Entrevista com T	11.8.2020 22:04	UCEWOonc	The Noite c	23	12.8.202
7	Vu6PNpYKu2U	DICAS DA RODAD	11.8.2020 19:14	UCJVbvkrBI	Cartoleiros	17	12.8.202
8	ly8jXKq_9AE	LIVE PLAYLIST DA	12.8.2020 5:31	UCg9nWuL	Tayara And	10	12.8.202
9	QAUqqcEU0Xc	PEDI ELA EM NAI	11.8.2020 2:02	UCOPS25A	PEIXE	24	12.8.202
10	eA4FRvf6vdM	AO VIVO - Apres	12.8.2020 2:58	UCZD5qcer	Vasco TV	17	12.8.202
11	8f70QZQB4UA	MASTERCHEF BR	12.8.2020 10:02	UC2EWGw	MasterChe	24	12.8.202
12	oH8wiqTGKrM	DIA DE FAZER CC	12.8.2020 1:36	UCIu-mBi1	PAMRIQUE	24	12.8.202
13	OxwD-3E6M-k	Kemilly Santos, A	11.8.2020 17:00	UCwS58Bc	KemillySan	10	12.8.202
14	uD5dJXCa_1s	Isadora Pompeo	11.8.2020 15:00	UCkskLrHR	Musile Rec	10	12.8.202
15	8irga_AqRdw	Minicurso Gratui	12.8.2020 4:16	UCeTKpYN	Gordices da	27	12.8.202
16	XZpj2Lx4HnA	REENCONTREI M	12.8.2020 0:54	UCp8i4boX	Jo?o Caetar	24	12.8.202
17	NQzNn_wQ_Vk	ESTOU LOIRA, DE	11.8.2020 21:08	UCmCEDd1	Thayna Tha	22	12.8.202
18	BTYfaXKDDHY	FREE FIRE AO VIV	11.8.2020 4:27	UCIVnGR9X	NFA CHANI	20	12.8.202
19	7WLxd6b2ayI	NÓS VOLTAMOS	11.8.2020 17:54	UCvym4Rx	Clone	24	12.8.202

Následne bolo veľmi problematické nastaviť manažéra pripájajúceho sa na CSV súbory v programe Microsoft SQL Server Data Tools. Problém s oddeľovačmi už bol síce vyriešený, ale bolo nutné správne nastaviť dátové typy tak, aby zbytočne nepreťažovali cieľovú databázu, ale aj obsiahli všetky potrebné dáta. Okrem toho tento manažér nebol schopný skonvertovať prečítané dátumy, takže musel byť zvolený dátový typ reťazec a nutnosť dodatočného spracovania v programe.

Obrázok 9: Nastavenie súborového manažéra v programe Microsoft SQL Server Data Tools, pre čítanie CSV súboru trending videí.



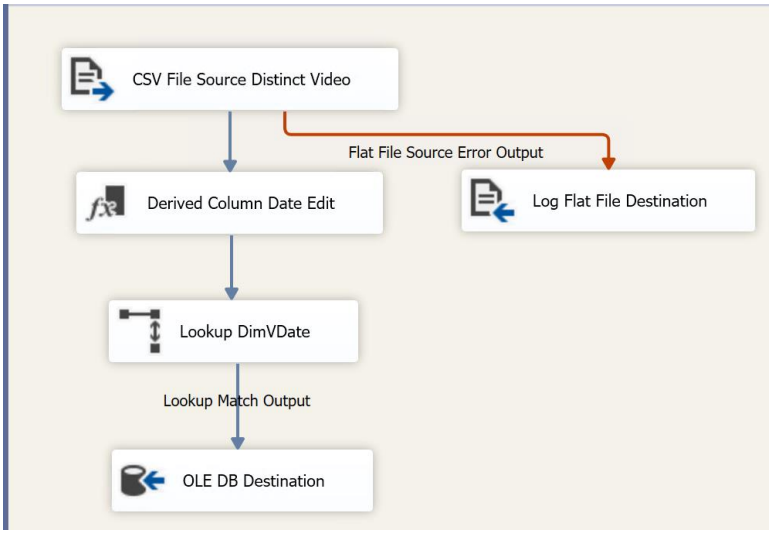
Po dostatočnom nastavení súborového manažéra, ktorý bol odladený za pomoci logovania chybových záznamov, bolo možné prejsť na úpravu dátumu. Dátum musel byť v požadovanom formáte a aby mohol byť pretypovaný a následne nájdený v časovej dimenzii. Pre tieto účely bol nižšie uvedený výraz použitý v Derived column, ktorý pridáva upravený dátum ako ďalší stĺpec k ostatným.

Obrázok 10: Výraz použitý v Derived column pre úpravu dátumov do požadovanej formy.

```
(DT_DBDATE)(SUBSTRING(TOKEN(publishedAt,".",3),1,4) + "-" +
RIGHT("0" + TOKEN(publishedAt,".",2),2) + "-" +
RIGHT("0" + TOKEN(publishedAt,".",1),2))
```

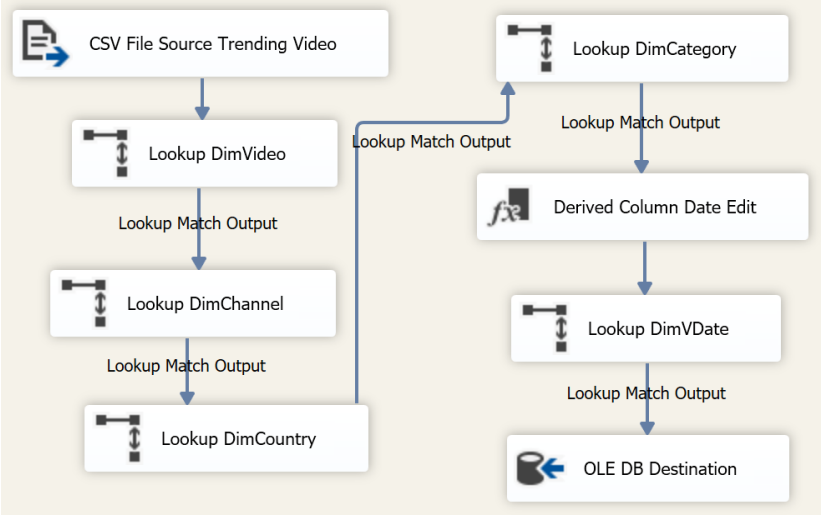
Získaný dátum je vyhľadaný v časovej dimenzii pod atribútom FullDateAlternateKey prostredníctvom prvku Lookup, ktorý vráti identifikačné číslo a nasleduje nahranie záznamov do dimenzionálnej tabuľky videí, ktorých bolo takmer 207 000.

Obrázok 11: SSIS nahrávajúci dáta do dimenzie video.



Predtým ako sa začali nahrávať dáta do faktickej tabuľky video, bolo nutné pripraviť dimenziu kanál. Tento proces je popísaný nižšie. S manažérom pre spracovanie faktických údajov už neboli ďalšie problému, bol potrebný nastaviť iba stĺpec pre krajinu.

Obrázok 12: SSIS nahrávajúci dáta do faktickej tabuľky video.



Prvým krokom po načítaní súboru bolo zistenie identifikačného čísla z dimenzie video a rovno tak aj pre dimenziu krajina a kategória. Po nájdení potrebných údajov, sa opäť upravil dátum kedy bolo video v trendoch a podľa neho získať identifikačné číslo. Výsledkom bolo v databáze 1,2 milióna záznamov.

Obrázok 13: SQL skript riešiaci problém duplicity vo faktickej tabuľke video.

```sql
insert into #Data (VideoKey, ChannelKey, CountryTrendingKey, TrendingDateKey)
SELECT [VideoKey]
      ,[ChannelKey]
      ,[CountryTrendingKey]
      ,[TrendingDateKey]
FROM [AW_DWH_xprchlik].[dbo].[factVideoDetail]
GROUP BY [VideoKey]
      ,[ChannelKey]
      ,[CountryTrendingKey]
      ,[TrendingDateKey]
HAVING COUNT(*) > 1


delete from [AW_DWH_xprchlik].[dbo].[factVideoDetail]
where VideoKey in (SELECT [VideoKey] FROM #Data)
and [ChannelKey] in (SELECT [ChannelKey] FROM #Data)
and [CountryTrendingKey] in (SELECT [CountryTrendingKey] FROM #Data)
and [TrendingDateKey] in (SELECT [TrendingDateKey] FROM #Data)
```

Pri analýze sa zistili nedostatky v dátach, kde ešte bolo nutné odstrániť duplicity a to za pomoci zložitejšieho mazacieho SQL skriptu, čím sa odstránilo približne 50 000 záznamov.

3.6. Spracovanie dát pre dimenziu kanál

Hlavným problémom pri získavaní kanálov pre dimenziu v dátasete o 1000 najlepších YouTuberoch bola absencia identifikačného kódu kanálu, podľa ktorého by bolo možné párovať videá s kanálmi. Našťastie bol k dispozícií url odkaz na kanál, ktorý vo svojej ceste obsahoval tento kód. Jednoduchým skopírovaním stĺpca v Excele a nahradením nadbytočnej časti prázdnym reťazcom, bol získaný identifikátor.

Obrázok 14: CSV súbor datasetu Top 1000 youtuberov otvorený v progrem Microsoft Excel.ä

	A	B	C	
1	Rank	username	Youtube Url	Name
2	1	tseries	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy4_brA	T-Series
3	2	checkgate	http://youtube.com/channel/UCbCmjCuTUZos6Inko4u57UQ	Cocomelon - Nurs
4	3	PewDiePie	http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_AJ5Yw	PewDiePie
5	4	MrBeast6000	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8uQQuVA	MrBeast
6	5	✿ Kids Diana Show	http://youtube.com/channel/UCk8GzjMOrta8yxDcKfylJYw	✿ Kids Diana Show
7	6	Like Nastya	http://youtube.com/channel/UCJplp5SjeGSdVdwsfb9Q7lQ	Like Nastya
8	7	WWEFanNation	http://youtube.com/channel/UCJ5v_MCY6GNUBTO8-D3XoAg	WWE
9	8	zeemusiccompany	http://youtube.com/channel/UCFFbwnve3yF62-tVXkTyHqg	Zee Music Compa
10	9	Vlad and Niki	http://youtube.com/channel/UCvlE5gTbOvjiolFlEm-c_Ow	Vlad and Niki
11	10	5-Minute Crafts	http://youtube.com/channel/UC295-Dw_tDNtZXFeAPAW6Aw	5-Minute Crafts
12	11	GoldminesTelefilms	http://youtube.com/channel/UCyoXW-Dse7fURq30EWl_CUA	Goldmines
13	12	kidrauhl	http://youtube.com/channel/UCIwFjwMjI0y7PDBVEO9-bkQ	Justin Bieber
14	13	sabtv	http://youtube.com/channel/UC6-F5tO8uklgE9Zy8IvbdFw	Sony SAB
15	14	BANGTANTV	http://youtube.com/channel/UCLkAepWjdylmXSltofFvsYQ	BANGTANTV
16	15	ibighit	http://youtube.com/channel/UC3IZKseVpdzPSBaWxBxundA	HYBE LABELS
17	16	CanalKondZilla	http://youtube.com/channel/UCffDXn7ycAzwL2LDlbyWOTw	Canal KondZilla

Jasné však bolo, že 1000 kanálov nie je dostatočných pre obsiahnutie získaných trendových videí. Dostupný bol aj Dataset kanálov, ktorý obsahuje množstvo záznamov,

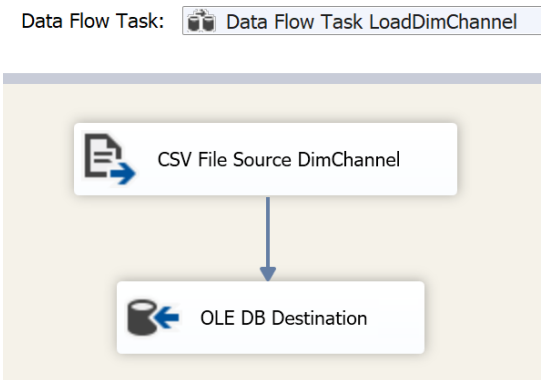
avšak okrem identifikačného kódu nemal žiadne popisné dáta, ktoré by bolo možné použiť v dimenzii. Preto boli kanály vytiahnuté aj zo súborov o trendových videách, ktorý mal v sebe ako identifikátory kanálov, tak aj ich názov a url adresa bola jednoducho doplniteľná.

Obrázok 15: Popisné dáta pre dimenziu video, nachádzajúce sa v datasete trending videí.

C	D	E	F
publishedAt	channelId	channelTitle	categoryId
12.8.2020 0:21	UCGfBwrCoi9ZJjKiUK8MmJNw	Pietro Guedes	22
11.8.2020 17:00	UCaO6TYtlC8U5ttz62hTrZgg	JYP Entertainment	10
10.8.2020 16:59	UCoXZmVma073v5G1cW82UKkA	As Irmãs Mota	22
11.8.2020 17:00	UC6Xqz2pm50gDCORYztqhDpg	League of Legends BR	20
11.8.2020 22:04	UCEWOoncsrmirqnFqxer9lmA	The Noite com Danilo Gentili	23
11.8.2020 19:14	UCJVbvkrBLp7L2pnaqc5CmQQ	Cartoleiros Gazeta do Povo	17
12.8.2020 5:31	UCg9nWuUlSG69Hv2VaCrE72w	Tayara Andreza	10
11.8.2020 2:02	UCOPS25AxMB9te9_-AHt3JEg	PEIXE	24
12.8.2020 2:58	UCZD5qcen7lbLPFTjfvdLFcw	Vasco TV	17
12.8.2020 10:02	UC2EWGw-KBjEReUbXMJEiaCA	MasterChef Brasil	24
12.8.2020 1:36	UClu-mBi1wc4Dt-WPzR0xRvA	PAMRIQUE	24
11.8.2020 17:00	UCwS58BcJEKW5huj_ZXESBww	KemillySantosVEVO	10
11.8.2020 15:00	UCkskLrHR3ga1AG_QS-trE6w	Musile Records	10
12.8.2020 4:16	UCeTKpYNnUeJ3g_9pbCpt3XA	Gordices da Deia	27

Dáta boli očistené od duplicít a za pomoci jednoduchého SSIS procesu nahraná do databáze. Získaných bolo skoro 25 000 kanálov.

Obrázok 16: SSIS nahrávajúci dáta do dimenzie kanál.



3.7. Spracovanie dát pre faktickú tabuľku kanálov

Dataset s 1000 najlepšími YouTubermi skrýval nemilé prekvapenie v spôsobe uloženia dát. Vo vlastnostiach, v ktorých by bol bežne očakávaný numerický dátový typ, boli dáta uložená v textovom dátovom typu reťazec. Po ručnom prehliadnutí dát bolo zistené, že čísla pre počet odberateľov, priemerný počet pozretí, priemerný počet palcov hore a priemerný počet komentárov sú uložené v skratke. Pro označenie miliónov bola použitá značka M a pre označenie tisícov naopak značka K.

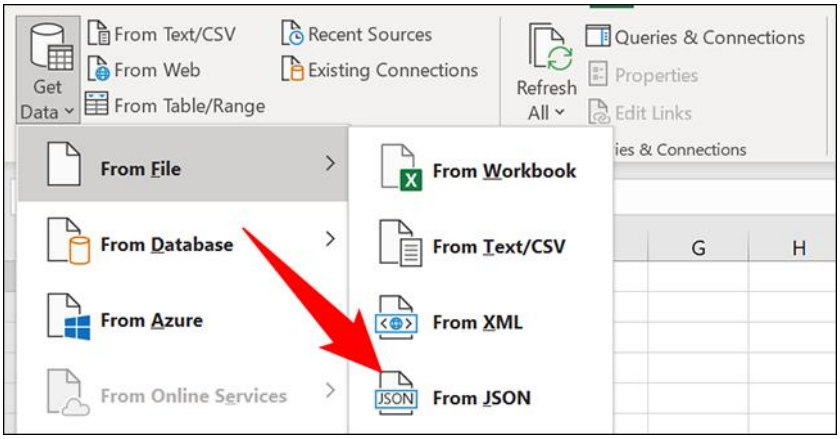
Obrázok 17: Ukážka nevhodných dát, ktoré bolo nutné ošetriť.

Avg, Views	Avg, Likes	Avg, Comments	Subscribers2
52,7K	1,7K	108	220100000
14,5M	79,5K		138600000
3,6M	244K	11,4K	111400000
36,7M	2,1M	78,5K	98400000
18,1M	71,8K		97500000
9,3M	52,6K		97500000
202,3K	6,6K	307	89400000
120,7K	4,9K	166	85700000
3,8M	17,7K		83600000
190K	1,8K	84	77100000

Úprava týchto dát prebehla v programu Microsoft Excel za pomoci funkcií "Hľadať" a "Dosadiť". Pomocou funkcie hladať bolo zistené, či sa v konkrétnej hodnote nachádza desetinná čárka, či se jedná o celé číslo. Funkcia Dosadiť nahradzuje konkrétny zadaný znak iným zvoleným znakom. Táto funkcia bola použitá trikrát – najskôr sa jednalo o odstránenie desatinnej čiarky (teda výmena desatinnej čiarky za prázdny reťazec). Druhé a tretie použitie nahradilo značky K a M potrebným počtom núl na základe práve tejto značky a práve prítomnosti alebo absencie desatinnej čiarky.

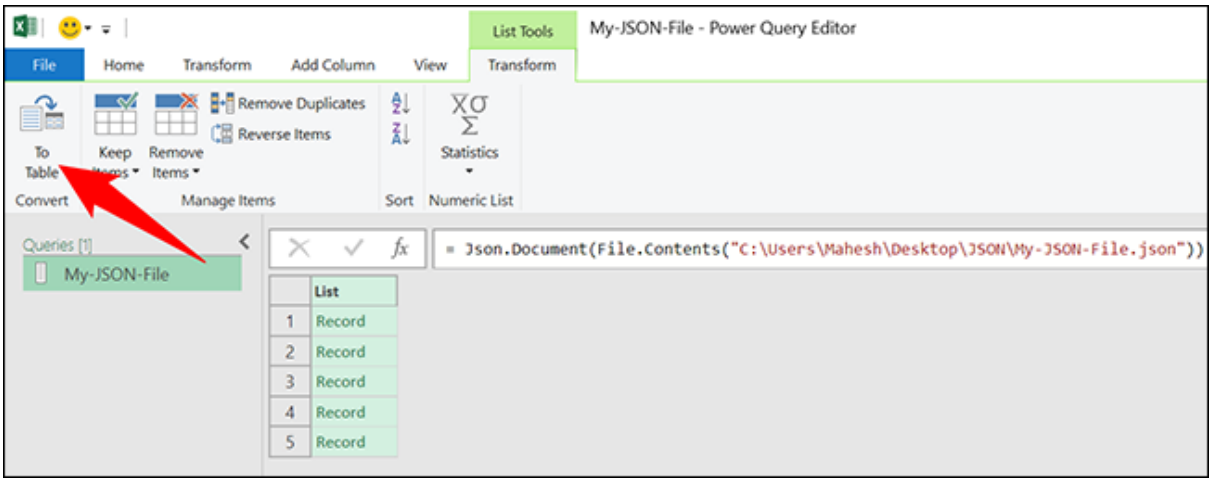
Týchto pár číselných údajov bolo veľmi nedostačujúcich a preto bol pripravený práve JSON súbor obsahujúci množstvo dát o videách a kanáloch, ako boli počty videí, celkový počet palcov hore a dole, komentárov, pozretí a stráveného času pozeraním kanálu užívateľmi. Pre otvorenie JSON súboru v Microsoft Excel bol použitý nasledujúci návod popísaný pomocou obrázkov.

Obrázok 18: Prvý krok pri otváraní JSON súboru v programe Microsoft Excel.



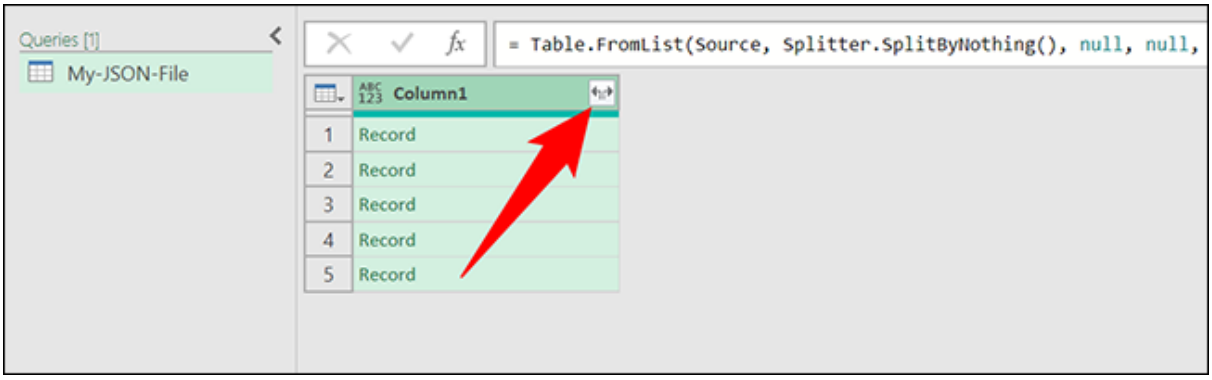
Najskôr sa súbor otvoril podobne ako CSV súbory, po dlhom načítavaní bol list záznamov prekonvertovaný na tabuľku.

Obrázok 19: Po prečítaní JSON súboru, je získaný list záznamov prekonvertovaný na tabuľku.

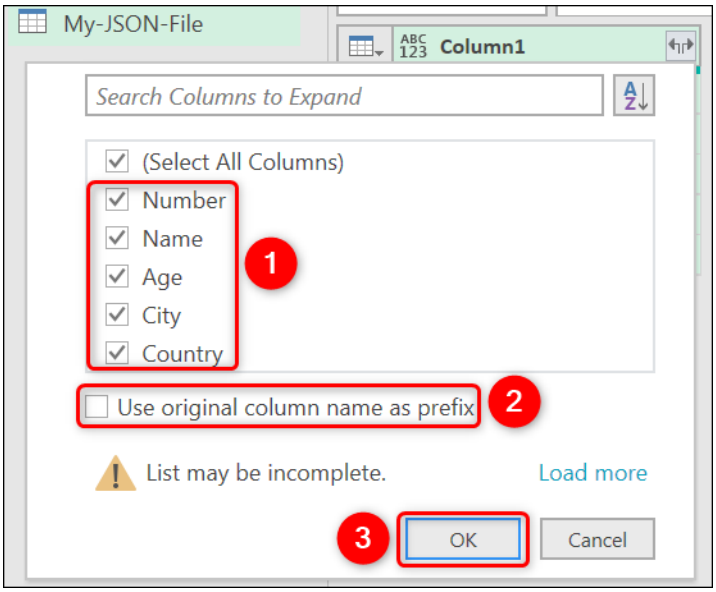


Výsledkom však bol iba list s jedným stĺpcom, kde sa po rozkliknutí ikony zobrazilo menu s dostupnými vlastnosťami dát.

Obrázok 20: Získaná tabuľka s komplexným stĺpcom obsahujúcim všetky atribúty záznamu.

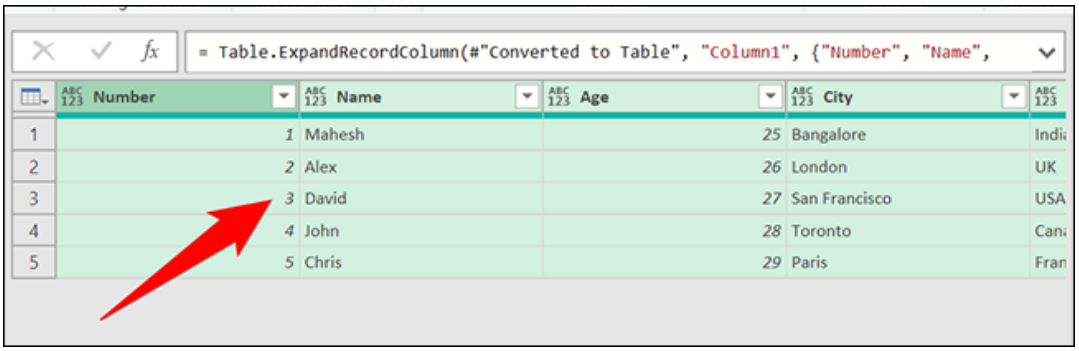


Obrázok 21: Rozdelenie komplexného stĺpca na požadovaná atribúty ako stĺpce.



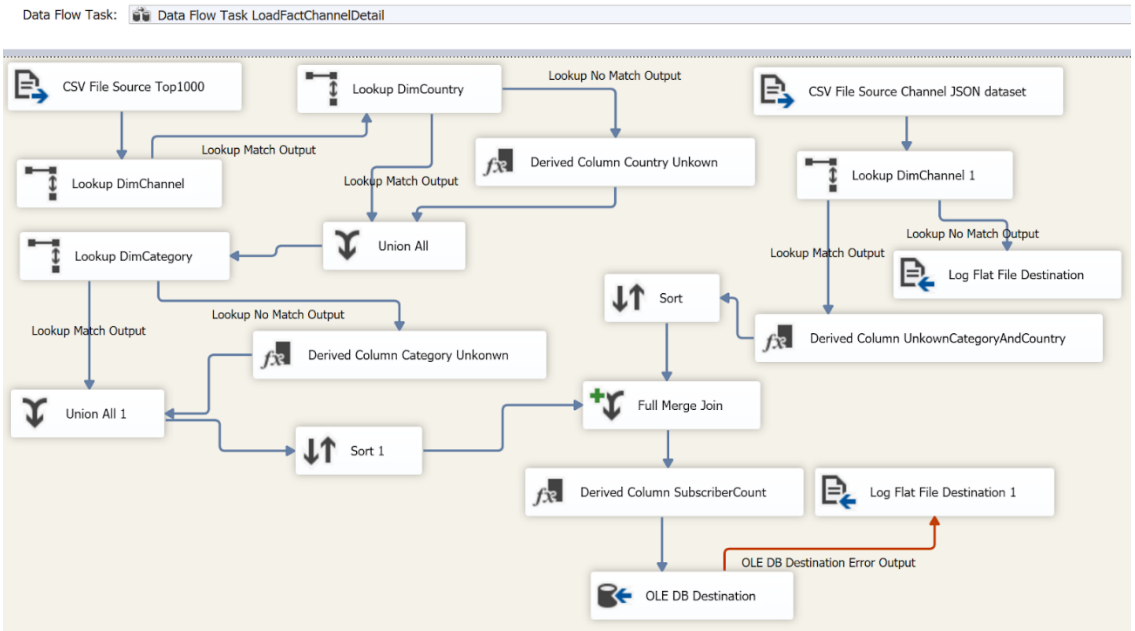
Zvolením požadovaných atribútov bola docielená konečná tabuľka, ktorá bola následne uložená do CSV súboru a prehnaná SSIS procesom.

Obrázok 22: Finálna tabuľka pripravená pre potrebnú transformáciu.



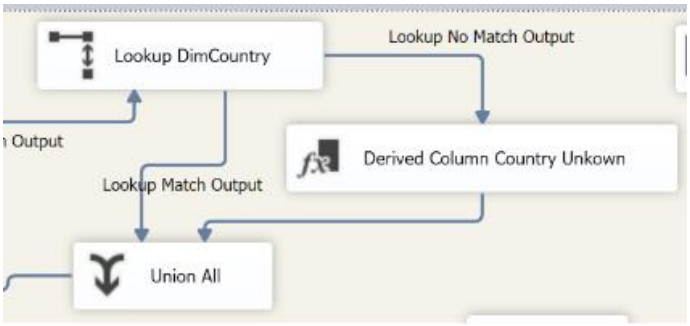
Zmienený SSIS proces, pre spracovanie údajov určených faktovej tabuľke kanálov, bol najzložitejšie, ktoré boli v rámci tohto projektu vytvorený. Bolo potrebné čítať dva CSV súbory zároveň, vykonať s ich záznamami nevyhnutné úkony, následne výsledky spojiť a nahrať do databáze.

Obrázok 23: Komplexný SSIS načítavajúci dáta z dvoch zdrojov pre faktickú tabuľku kanál.



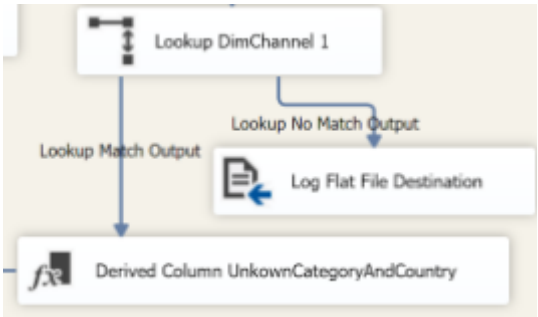
Primárnym súborom, ktorý obsahoval aj údaje o hlavnej krajine publika a kategórie aké videá nahráva, bol ten o 1000 najlepších YouTuberoch. Keďže obsahoval tieto údaje, využívalo sa vyhľadávanie v príslušných dimenziach. Špecifikum bolo, že ak sa nenašla zhoda, alebo údaj chýbal, záznam sa presmeroval do bloku upravujúceho chybné hodnoty a nastavil kľúče na neznámu hodnotu. Toto rozdelenie bolo potrebné opäť zjednotiť a pokračovať v práci ako je ukázané nižšie.

Obrázok 24: Postup pri získavaní identifikačného čísla krajiny a ošetrenie záznamov s nenájdenými ID.



V prípade JSON dátasetu bol veľký problém v tom, že obrovské množstvo dát nenašlo zhodu v dimenzí kanálov, čo projekt do istej miery obmedzil pri analýze a možných výsledkoch. Avšak bolo potrebné ďalej pracovať s tým čo bolo k dispozícii. Po nájdení zhody s kanálom, boli pridané iba dva chýbajúce stĺpce o krajine a kategórii kanálu a označená za neznáme.

Obrázok 25: Vylúčenie kanálov z JSON súboru pri nenájdení zhody a ošetrenie chýbajúcich atribútov.



Dokončením práce na dopĺňaní informácií bolo nutné záznamy zoradiť a plným spojením podľa identifikačného čísla kanálu zlúčiť. Po zlúčení dát už zostávalo iba vysporiadať sa s duplicitnou informáciou o počte odberateľov kanálu. Ako prioritná hodnota sa brala od súboru s 1000 najlepšími a to z dôvodu aktuálnejších dát. Ako je možné vidieť nižšie, boli prázdne hodnoty boli nahradené číslom 0. Následne boli podmienkou zistené nulové hodnoty a nahradené údajom z JSON súboru. Do databáze bolo nahraných takmer 4000 záznamov.

Obrázok 26: Ošetrenie chýbajúcich a neaktuálnych údajov odberateľov kanálu.ä

```
(REPLACENULL(SubscriberCount,0) == 0)
? SubscriberCount2 : SubscriberCount
```

3.8. Spracovanie dát pre dimenziu a faktovú tabuľku značka

Dáta pre dimenzionálnu a faktovú tabuľku značiek videí pochádzajú zo súborov o trendových videách. Pre spracovanie sa vyňali stĺpce o identifikačných kódov videí a značky pridelené videu. Toto vyňatie bolo uskutočnené až po základných úpravách týchto súborov, ktoré sú spomenuté v sekcí spracovania dát pre dimenziu a faktovú tabuľku videa. Značky videa, ak

boli prítomné, boli vo forme reťazca oddelené znakom |. Túto formu bolo nutné rozložiť a to do podoby, jeden identifikačný kód a jedna značka.

Obrázok 27: Vyextrahované značky videí z datasetu trending videí.

	A	B	C
1	video_id	tags	
2	s9FH4rDMvds	pietro	
3	s9FH4rDMvds	guedes	
4	s9FH4rDMvds	ingrid	
5	s9FH4rDMvds	ohara	
6	s9FH4rDMvds	pingrid	
7	s9FH4rDMvds	vlog	
8	s9FH4rDMvds	amigos	
9	s9FH4rDMvds	jully	
10	s9FH4rDMvds	molina	

Bol vytvorený skript v programovacom jazyku Python, ktorý v príkazovom riadku prijal vstup s názvom spracovávaného súboru. Následne na základe oddeľovača ; pre CSV, boli rozdelené identifikátory a značky. Odstránili sa prázdne hodnoty a ďalej sa rozdelil reťazec značiek. Výsledok bol zapísaný do CSV súboru. Celý kód je možné vidieť nižšie.

Obrázok 28: Skript v programovacom jazyku Python pre spracovanie značiek videí.

```
1   import sys
2   import csv
3
4   fixed_stdin = map(lambda x: x.replace("\r", " ").replace("\0", ""), sys.stdin)
5   writer = csv.writer(sys.stdout, delimiter=",")
6   name = None
7   tags_str = None
8   try:
9       for name, tags_str in csv.reader(fixed_stdin, delimiter=";"):
10          if tags_str != '[None]':
11              for tags in csv.reader((tags_str,), delimiter="|"):
12                  for tag in tags:
13                      writer.writerow((name, tag))
14  except Exception as e:
15      print("chybne tagy:", name, repr(tags_str), file=sys.stderr)
16      raise e
```

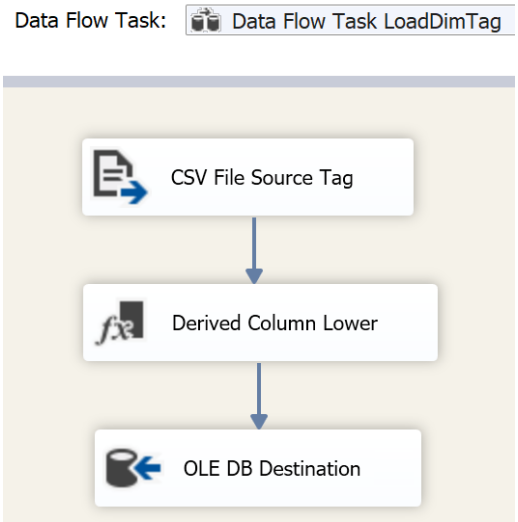
Pre potreby dimenzie bol vytvorený CSV súbor, ktorý obsahoval všetky značky, očistené od duplicít pomocou programu Microsoft Excel. Rovnako bol očistený aj súbor s identifikátormi a značkami, ale s rozdielom, že sa hľadali duplicity podľa oboch vlastností. Bohužiaľ ani to nakoniec nestačilo a bolo nutné vykonať zásah až v databáze SQL výrazom, ktorý je nižšie.

Obrázok 29: SQL skript pre vymazanie duplicitných záznamov v bezfakticko faktickej tabuľke značiek.

```sql
delete from [dbo].[factTag]
    where VideoKey in (SELECT [VideoKey]
        FROM [dbo].[factTag]
        GROUP BY [VideoKey], [TagKey]
        HAVING COUNT(*) > 1)
    and TagKey in (SELECT [TagKey]
        FROM [dbo].[factTag]
        GROUP BY [VideoKey], [TagKey]
        HAVING COUNT(*) > 1)
```

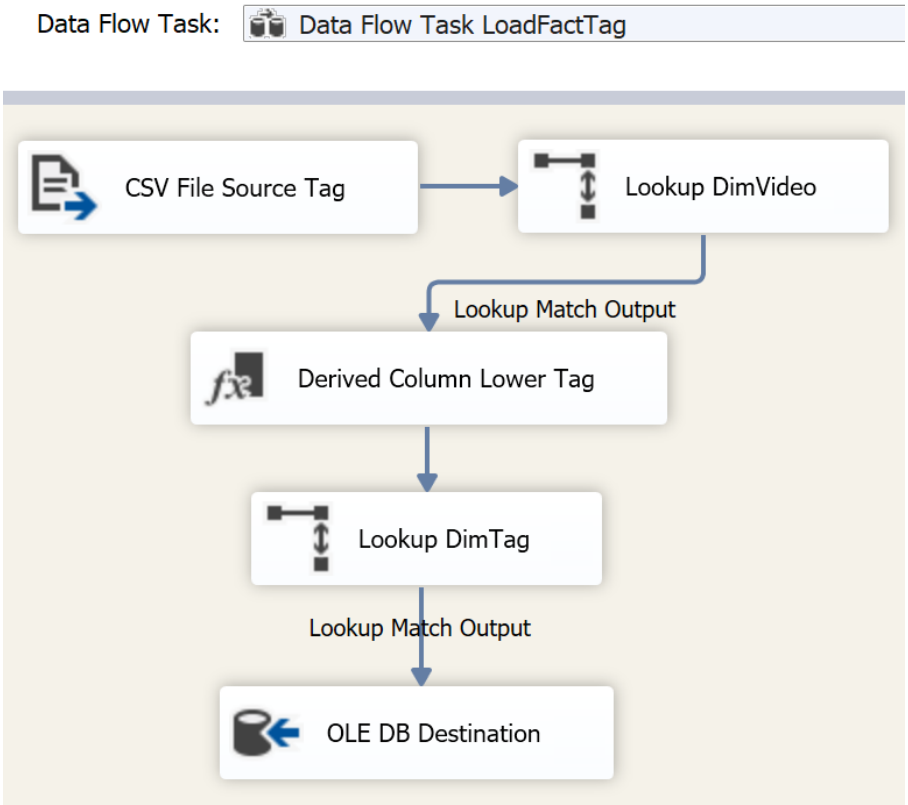
Nahranie do dimenzionálnej tabuľky programom Microsoft SQL Server Data Tools, bolo jednoduché. Nutnosťou bolo iba zmeniť značku na malé písmená, kvôli náchádzaniam zhôd vo faktickej tabuľke. Zaznamenaných bolo 850 000 značiek.

Obrázok 30: SSIS nahrávajúci dáta do dimenzie značka.



Spomínaná faktická tabuľka nie je klasická. Ako už bolo spomenuté v kapitole štruktúra SQL databáze, jedná sa o takzvanú bezfaktovú faktovú tabuľku, ktorá predstavuje mostové spojenie medzi dimenziami video a značka. Toto spojenie je nutné kvôli väzbe veľa k veľa, kde video môže mať viac značiek a značka môže patriť k viacerým videám.

Obrázok 31: SSIS nahrávajúci dáta do bezfaktickej faktickej tabuľky značka.



Nahranie do databáze je zložené z viacerých prvkov. Zistenie identifikačného čísla videa na základe identifikačného kódu pri značke, následne zmenenie znakov značky na malé písmená a podľa nich vyhľadanie identifikátoru v dimenzí značka. Po zmienenom odstránení duplicít v databáze, ostalo takmer 2 milióny záznamov o značkách.

4. Predpríprava vizualizácie dát

Pre analýzu a vizualizáciu dát bola použitá platforma Power BI, v ktorej bolo vytvorené niekoľko reportov zobrazujúce naše predom definované KPI.

4.1. Nahranie a príprava dát do Power BI

Dáta boli najprv nahrané do Power Bi pomocou možnosti Údaje – SQL server – Databáza SQL Serveru, kde sme sa pripojili na daný server treeman.mendelu.cz. Po pridaní dát do Power BI sme narazili na pár problémov, ktorých riešenie predchádzalo niekoľko krokov:

Prvým krokom bola úprava zdroja pred nahratím do Power BI. V tomto kroku sme sa rozhodli vytvoriť si ku každej faktickej tabuľke pohľad (ďalej view) na dáta. Toto riešenie pri analýze a vizualizácií dát uľahčuje prácu, je zrozumiteľnejšie nie len pre užívateľa ale aj pre zákazníka, a to hlavne preto, pretože sú v tabuľkách zobrazené len slová bez zbytočných id a kľúčov. Ďalšou výhodou pre nás bolo aj to, že sa nám predtým v niektorých prípadoch vytváral mostík pri filtrovaní (napríklad kategória pre Video a Kanál). Vytvorenie týchto view bolo urobené v Microsoft Server SQL štúdiu a vyzeralo následne:

Obrázok 32: Script pre vytvorenie ViewVideo

```
create view [viewVideoDetail] as
SELECT  [VideoDetailKey]
       ,[VideoKey]
     ,dch.ChannelTitle
     ,dch.ChannelId,
     dch.ChannelKey,
     dch.Username,
     dch.YoutubeUrl
      ,f.CountryTrendingKey
      ,dco.CountryName,
      dco.Code as country_code
      ,f.CategoryKey
      ,dca.CategoryName,
      dca.CategoryTypeKey
      , dcat.CategoryTypeName
      ,format(d.FullDateAlternateKey,'yyyy-MM') as monthdate
      ,d.FullDateAlternateKey
      ,TrendingDateKey
      ,[ViewCount]
      ,[Likes]
      ,[Dislikes]
      ,[CommentCount]
   FROM [AW_DWH_xprchlik].[dbo].[factVideoDetail] f
  left join dimChannel dch
    on f.ChannelKey=dch.ChannelKey
 left join dimCountry dco
 on f.CountryTrendingKey = dco.CuontryKey
 left join dimCategory dca
 on dca.CategoryKey = f.CategoryKey
 left join dimCategoryType dcat
on dca.CategoryTypeKey = dcat.CategoryTypeKey
  left join [AW_DWH_xprchlik].[dbo].dimVDate d
    on f.TrendingDateKey = d.DateKey
    ;
```

Obrázok 33: Script pre vytvorenie ViewKanál

```
create view ViewChannelDetail AS
SELECT [ChannelDetailKey]
      ,[ChannelKey]
      ,[CountryAaudienceKey]
      ,dc.CountryName
      ,dc.Code as country_code
      ,f.CategoryKey
      ,dca.CategoryName
      ,dca.CategoryID
      ,dcat.CategoryTypeName
      ,dcat.CategoryTypeKey
      ,[Rank]
      ,[ElapsedTime]
      ,[ViewCount]
      ,[VideoCount]
      ,[CommentCount]
      ,[SubscriberCount]
      ,[AvgViews]
      ,[AvgLikes]
      ,[AvgComments]
  FROM [AW_DWH_xprchlik].[dbo].[factChannelDetail] f
  LEFT JOIN dimCountry dc
    ON f.CountryAaudienceKey = dc.CuontryKey
  LEFT JOIN dimCategory dca
    ON f.CategoryKey = dca.CategoryKey
  LEFT JOIN dimCategoryType dcat
    ON dca.CategoryTypeKey = dcat.CategoryTypeKey;
```

Toto riešenie nebolo najefektívnejšie ale vzhľadom na čas, ktorý nás tlačil nám prišiel ako najvhodnejší. V prípadnej budúcej úprave by bolo vhodnejšie vytvoriť transformáciu, ktorá by bola tiež automatizovaná a vytvárala by rovno tabuľky vo forme týchto view, ktoré by boli jednoducho aktualizovateľnejšie.

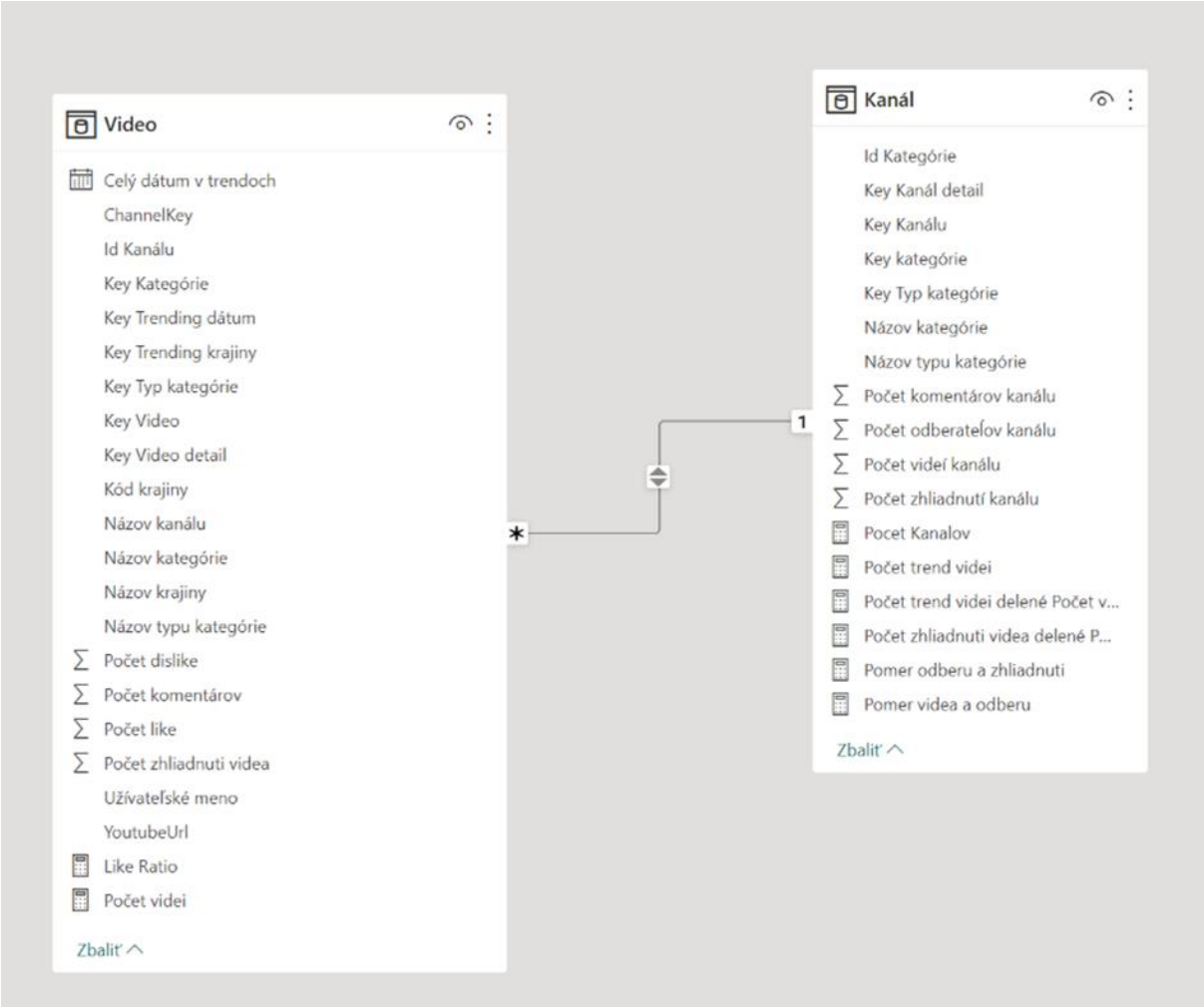
Ďalej sme si tieto vytvorené view exportovali a to v Management štúdiu pomocou: export file – zvolíme si názov databáze – tasks – export data – zvolim source – a výstup zvolím flatfile. Súbor sme si exportovali aby sme mohli pracovať na vizualizácií dát aj z domu. Následne sme v Power BI nahrali dáta. Tento krát sme zvolili ako zdroj dát možnosť Text/CSV.

Po nahratí dát sme narazili na ďalší problém a tým bolo to, že medzi tabuľkami vznikol v Power BI vzťah m x n. K tomuto vzťahu však nedošlo z dôvodu chybnej analýzy dát a následnému zlému návrhu dimenzionálneho modelu, ale kvôli chybe v načítavaní dát za pomoci SSIS procesu. Pri nahrávaní dát z dvoch zdrojov o kanáloch, bol použitý prvok zjednocujúci záznamy a teda vznikli duplicity v dátach. Tento problém bol vyriešený úpravou tohto procosu s využitím zlučovacieho prvku na základe identifikačných čísel kanálov. Okrem týchto problémov sme boli nútený z analýzy dát vyradiť číselné údaje o kanáloch pochádzajúcich zo súboru o najlepších 1000 Youtuberoch. Nebol totiž jasný ich význam, kvôli absencii popisu na stránkach datasetu.

4.2. Úprava dát v Power BI

Po úspešnom nahratí dát a prepojení faktických tabuliek v Modelu, zobrazenému na obrázku nižšie, sme upravili názvy atribútov do zrozumiteľnejšej podoby.

Obrázok 34: Model v PowerBI



Ďalej sme upravili v tabuľkách dátové typy atribútov a naformátovali sme si zostavu (úpravu písma, farieb a pod.) pre vizualizáciu dát.

5. Výsledky

V prvom kroku bola v DAXE vytvorená nová metrika pre uľahčenie výpočtov týkajúce sa našich KPI.

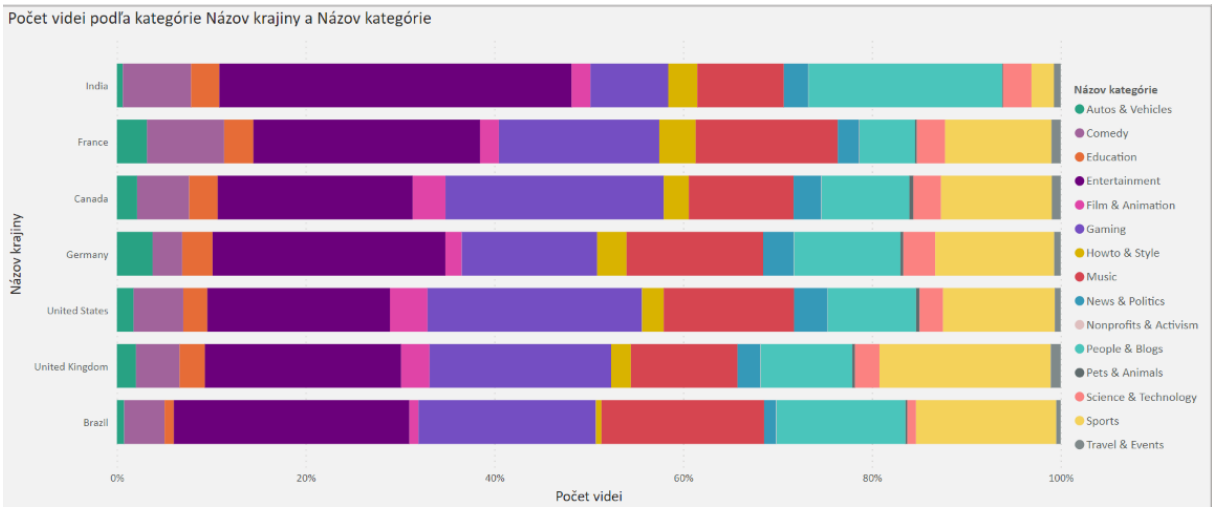
```
Počet videi = DISTINCTCOUNT(Video[Key Video])
```

Tento počet vracia unikátny počet videí vyskytujúcich sa v trendoch.

5.1. Obľúbenosť kategórií videa v závislosti od krajiny

Pre zobrazenie popularity - obsadenia jednotlivých kategórií videa bol využitý skladový graf. Na ktorého x osy je nanesený unikátny počet videi, a na ose y názov krajiny. Legenda nám zobrazuje názov kategórie videa.

Obrázok 35: Počet videí podľa kategórie videa v krajine



Ako môžeme vidieť na obrázku vyššie, najviac prevládajúca kategória vo všetkých krajinách je kategória zábava. Môžeme vidieť, že najväčšie zastúpenie tejto kategórie je v Indii, a to presne v 41%. Ďalšie zastúpenia v krajinách: Nemecko (25,7%), Brazília (25,4%), .., USA (19,6%). Môžeme teda povedať, že naša **hypotéza: „Má kategória videí viac ako 30% zastúpenie v každej krajine“ je nepotvrdená.**

Ďalej sme sa zamerali na najväčšie zastúpenie jednotlivých kategórií. Najviac obsadená kategória videí v trendoch o autách je v Nemecku. Najväčšie obsadenie komediálnych videí v trendoch je vo Francúzsku. Trendové videa označené kategóriou hry prevažuje v USA. Športové trendové videá prevažujú v Anglicku a videá v trendoch zamerané na hudbu v Brazílii. Naopak najmenšie obsadenie v trendoch pre kategóriu autá a šport je zastúpená v Indii, kategória komédia v Nemecku, kategória veda a technika v Brazílii.

5.2. Počet trendových videí pre kanál pre každý mesiac/rok

Dopočítaná metrika v DAXE:

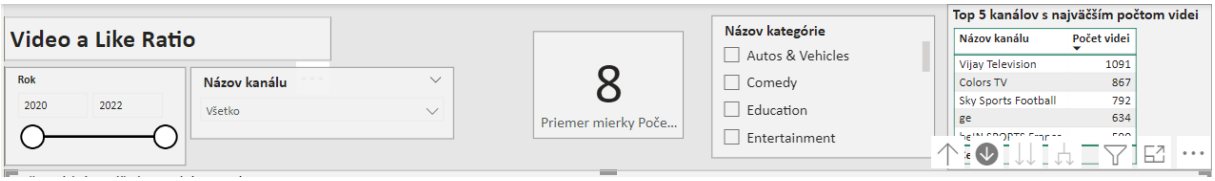
```
Priemer mierky Počet videi na kategóriu Celý dátum v trendoch
=
```

```
AVERAGEX(
    KEEPFILTERS(VALUES('Video'[Celý dátum v trendoch])),
    CALCULATE([Počet videi])
)
```

Pre zobrazenie výsledkov analýzy pre 2 a 3 KPI sme vytvorili niekoľko filtrov: podľa roku, názvu kanála a kategórie videa. Ďalej sa v reporte nachádza aj tabuľka s top 5 kanálmi s najväčším počtom trendových vydaných videí.

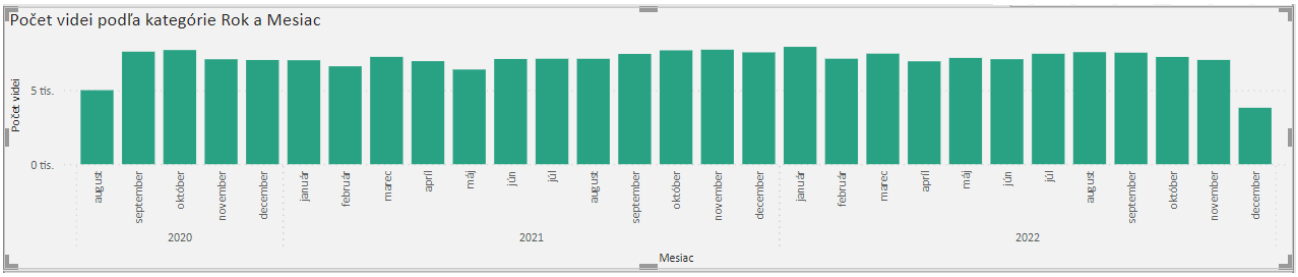
Ďalej môžeme v zapätí nájsť aj informáciu o priemernom počte trendových videí kanálu. Priemerný počet trendových videí na kanál je 8 a to celkovo za celé obdobie.

Obrázok 36: Zapätie pre 2 a 3 KPI

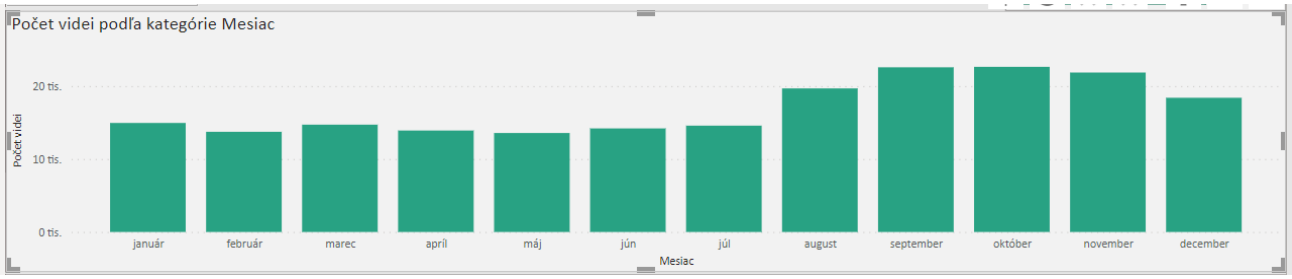


Ako môžeme vidieť na obrázku nižšie, na vytvorenie vizuálu sme využili skladaný stĺpcový graf. V grafe máme nanesené na ose y počet unikátnych videí a na ose x vytvorenú hierarchiu pre dátum (za rok alebo za rok a mesiac).

Obrázok 37: Počet trendových videí pre všetky kanále za každý mesiac a rok.



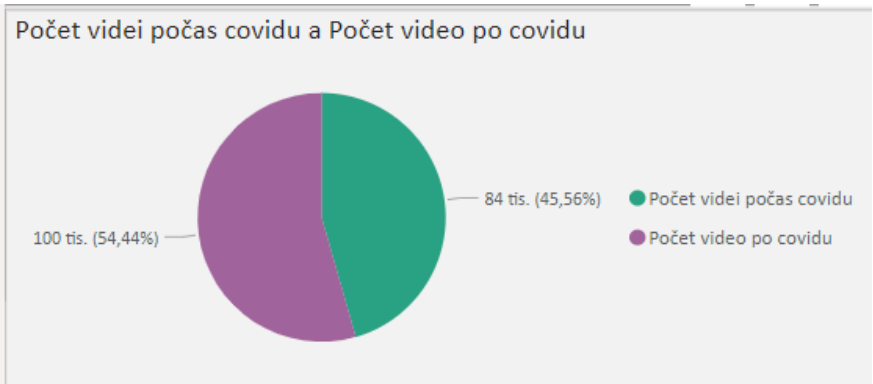
Obrázok 38: Počet trendových videí pre všetky kanále za každý mesiac.



Ako môžeme vidieť na obrázku vyššie, „úspešnými“ mesiacmi, v ktorých sa zvyšuje počet trendových videí sledovaných kanálov, sú mesiace: september, október a november. Je to možno spôsobené tým, že väčšina umelcov tvorí práve obsah v lete, a po lete ho vydáva.

Ďalej sme sa venovali našej predom stanovenej **hypotéze: „Počet videí na kanál sa počas koronavírusu zvýšil o viac ako 10%“**. Ako sme spomínali v úvode, koronavírusové obdobie berieme od počiatočného dátumu našich dát 7.2020 až po 5.2021, a po koronavírusové obdobie od 6.2021 po 12.2022. Jednotlivé počty pre obdobia sú zobrazené na obrázku nižšie.

Obrázok 39: Počet videí počas a po koronavírusovom období



Ďalším krokom bolo vypočítanie o koľko percent sa zmenil tento počet trendových videí medzi obdobiami. Novú metriku sme vypočítali pomocou DAX:

```
kolko percent = ([Počet video po covidu]/[Počet videi počas
covidu])-1
```

Tento výsledok sme si zobrazili v reporte a činí 0,19. **Hypotézu teda môžeme potvrdiť**. Počet trendových videí sa navýšil až o 19%.

5.3. Like ratio trendových videí pre kanál

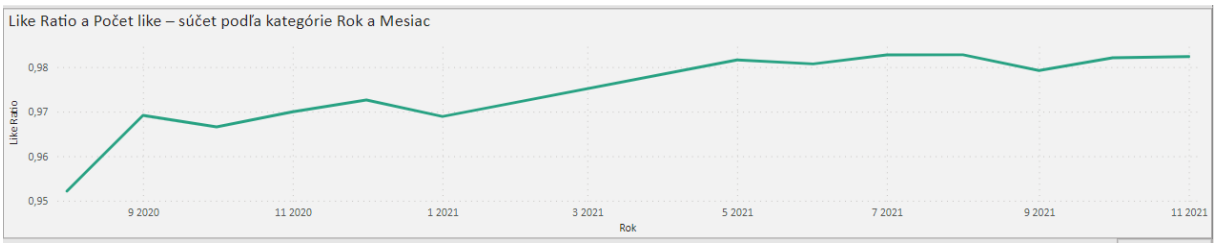
Dopočítaná metrika v DAXE:

```
Like Ratio = SUM(Video[Počet like])/(SUM(Video[Počet like])+
SUM(Video[Počet dislike]))
```

Ďalším sledovaným KPI je like ratio: pomer medzi počtom likov a celkovým počtom likov a dislikov. Podľa Zdroja Understanding The TikTok Algorithm – Ranieri Communications sa like ratio ešte považuje dobré, ak je vyššie ako 95 %. V našom prípade sledujeme, či video sa stalo trendovým kvôli tomu, či bolo viac likované, teda dobré, resp. bizarné.

Pre zobrazenie výsledkov je využitý čiarový graf. Na osy x je nanesená hierarchia dátumu a na osy y je nanesená dopočítané metrika – like ratio.

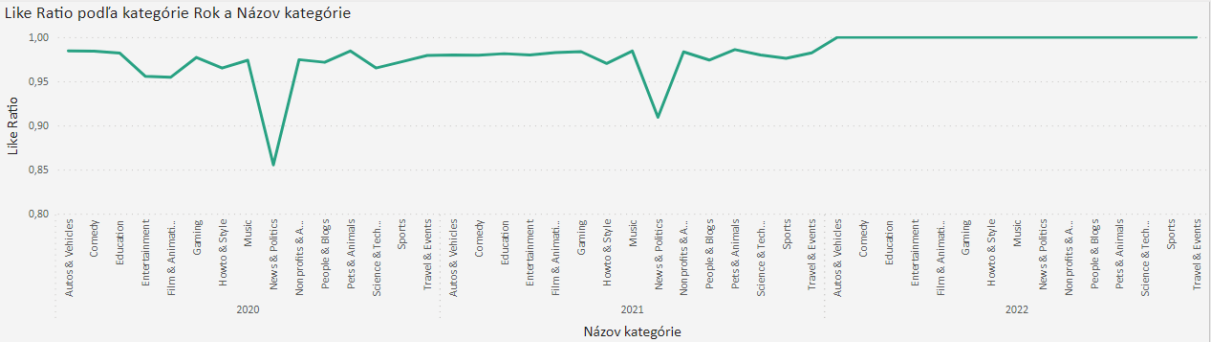
Obrázok 40: like ratio v čase pre kanál Vijay Television



Na obrázku vyššie môžeme vidieť ako sa like ratio menilo v čase pre kanál s najväčším počtom kanálov – pre kanál Vijay Television s 1091 videi, ktoré sa stali trendové. Môžeme si všimnúť, že za celý čas jeho pomer neklesol pod predom stanovených 95%.

Ďalej sme sa zamerali na kategórie, ktoré dosahujú nízke like ratio v rokoch. Ako môžeme vidieť na obrázku nižšie, jediná kategória ktorá ma nižšie ratio než 95%, v rokoch 2020 a 2021, je kategória novinky a politika. V roku 2020 sa pohybovalo okolo 86%, čo mohlo spôsobiť práve koronavírusové opatrenia, novinky súvisiace s koronavírusom, hoaxy a celkové rozdelenie spoločnosti.

Obrázok 41: Like ratio pre kategóriu v čase



V našich analyzovaných dátach – hodnoty like a dislike chýbali od roku 2022, preto sú výsledky len v období 2020 až 2021.

5.4. Pomer počtu trendových videí a celkového počtu videí pre obľúbený kanál

Dopočítané metriky v DAXE:

```
Počet trend videi = DISTINCTCOUNT('Video'[Key Video])
Počet trend videi delené Počet videí kanálu = DIVIDE([Počet
trend videi], SUM('Kanál'[Počet videí kanálu]))
```

Zobrazuje koľko percent zo všetkých videí obľúbeného kanálu tvoria videá ktoré sa ocitli v trendoch. Ako môžeme vidieť nižšie, bola vytvorená tabuľka, v ktorej sú zoradené obľúbené kanály podľa najvyššieho počtu trendových videí k celkovým.

Na základe vizualizácie je jednoznačne viditeľný trend v oceňovaní kvality videí, na rozdiel od kvantity. To je do istej miery veľmi prekvapivý jav v súčasnej konzumnej spoločnosti. Na opačnej strane pomeru nahraných videí kanálu a videí vyskytujúcich sa v trendoch je rovnako úplne obrátená. Dominujú kanáli s tisícami videí, kde sa mu náhodou dostalo jedno video do trendov. Nie je jednoznačne možné určiť dôvod tohto javu, ale často sa jedná o kanály hudobníkov alebo kanály venujúce sa ohurujúcim aktivitám alebo informáciám.

Ďalej môžeme vidieť na stránke aj priemerné obsadenie trendových videí k celkovým u všetkých kanálov.

Najvyššieho výsledku dosahuje kanál s názvom Kuzgesagt – In a Nutshell. Ktorého viac než 67 % videí sa ocitlo v trendoch, čo je až o 60 % viac než je priemerný počet u všetkých kanálov. Jedná sa o nemecký Youtube kanál, ktorého obsah tvoria krátke animované videá plné zaujímavostí najrôznejšieho druhu. Videá sú náučné, krátke a stručné, presne ako napovedá názov celého kanála. Užívatelia tu nájdu najrôznejšie fakty o vede – o vesmíre, sociálnych vedách, histórii, biológii. Podľa nášho názoru práve rozmanitosť, nápaditosť a hlavne jednoduché nevedecké podanie je kľúčom úspechu tohto kanálu.

Kanály s takmer nulovým pomerom sa neprekvapivo zaoberali správami alebo patrili k rôznym televíziám. To aj vysvetľuje množstvo nimi nahraných videí.

Obrázok 42: vizualizácia pomera počtu videí a celkového počtu videí kanálu

Pomer počtu trend videí a celkového počtu videí pre obľúbený kanál

7,02 %

Počet trend videi delené Počet videí kanálu

Názov kanálu	Počet trend videi	Počet videí kanálu – súčet	Počet trend videi delené Počet videí kanálu
Kurzgesagt – In a Nutshell	45	67	67,16 %
Dude Perfect	65	161	40,37 %
Henrique e Juliano	40	161	24,84 %
Veritasium	53	233	22,75 %
The Slow Mo Guys	28	156	17,95 %
The Game Theorists	68	426	15,96 %
twenty one pilots	11	82	13,41 %
SQUEEZIE	153	1144	13,37 %
Porta dos Fundos	94	740	12,70 %
Matt Stonie	35	292	11,99 %
JYP Entertainment	78	657	11,87 %
Coldplay	20	218	9,17 %
Shruti Arjun Anand	55	611	9,00 %
Jorge & Mateus Oficial	18	212	8,49 %
Hacksmith	41	498	8,23 %
Jake Paul	32	421	7,60 %
SSSniperWolf	118	1554	7,59 %
League of Legends	55	753	7,30 %
Vijay Television	1091	16031	6,81 %
Marques Brownlee	63	944	6,67 %
Shemaroo Comedy	49	796	6,16 %
Mrwhosetheboss	52	860	6,05 %
JJ Olatunji	47	801	5,87 %
Renato Garcia	66	1144	5,77 %
White Hill Music	57	1005	5,67 %
MyMissAnand	4	77	5,19 %
NikkieTutorials	32	626	5,11 %
Times Music	37	815	4,54 %
Manual do Mundo	52	1244	4,18 %
EU FICO LOKO	28	670	4,18 %
DisneyMusicVEVO	21	503	4,17 %
Sony SAB	589	14151	4,16 %
Miniminter	54	1428	3,78 %
Unbox Therapy	48	1274	3,77 %
Sony Music India	70	1999	3,50 %
Celkovo	**7656**	**1513353**	**0,51 %**

5.5. Pomer počtu videí a odberov pre obľúbený kanál

Dopočítaná metrika v DAXE:

```
Pomer videa a odberu = SUM('Kanál'[Počet videí
kanálu])/SUM('Kanál'[Počet odberateĺov kanálu])
```

Ďalším sledovaným KPI je pomer počtu vydaných videí a odberateľov pre obľúbený kanál. Toto KPI ukazuje, či kanál vydáva veľa videí a má veľa odberateľov – tým nižšie číslo, alebo vydáva veľa videí ale má málo odberateľov – tým väčšie číslo.

Ako môžeme vidieť na obrázku nižšie, mnoho odberateľov a zároveň málo videí vydáva kanál Dude Perfect, ktorého pomer je veľmi nízky. Naopak najviac videí ale zároveň málo odberateľov vydáva kanál CNN, ktorý sa od priemeru 0,00014 odchyľuje až o 0,0097. Tieto údaje plne zodpovedajú výsledkom sledovania pomeru trending videí a celkových videí.

Ďalej môžeme vidieť v reporte, že najvyššie čísla dosahujú kategórie: Politika, Hudba, Šport. Najnižšie miesto, ktoré CNN zaujalo, nás veľmi neprekvapilo. Mnoho ľudí sa pozrie na zaujímavé správy, málokto z nich však cielene kanál CNN sleduje. Sledovanie takéhoto kanála väčšinou indikuje záujem o väčšinu videí, ktoré kanál produkuje – v takomto množstve videí najrôznejších žánrov je však odber veľmi nepravdepodobný.

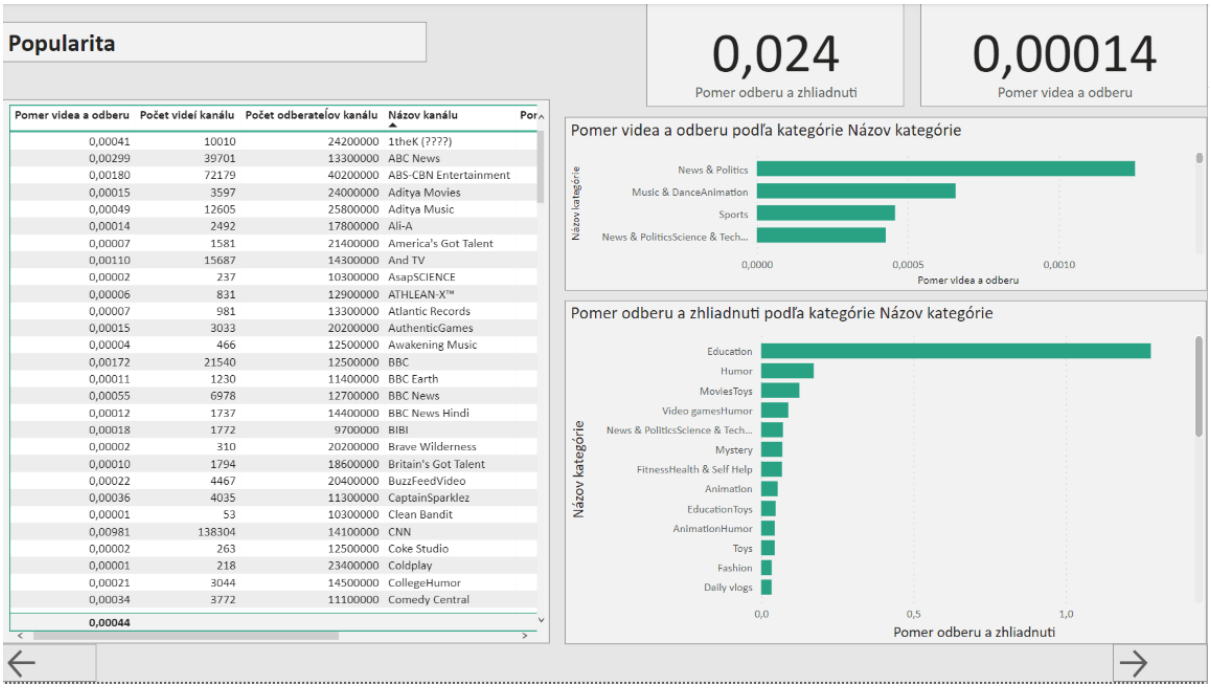
5.6. Pomer počtu odberov a zhliadnutie obľúbeného kanálu

Dopočítaná metrika v DAXE:

```
Pomer odberu a zhliadnuti = SUM('Kanál'[Počet odberateĺov
kanálu])/SUM('Kanál'[Počet zhliadnutí kanálu])
```

Posledné sledované KPI nám zobrazuje pomer medzi počtom odberateľov a celkovým počtom vzhliadnutí obľúbeného kanálu. Na obrázku nižšie môžeme vidieť, že najviac sa od priemeru – 0,024 odchyľuje kategória Vzdelávanie (1, 28) od ostatných kategórií. Čo znamená, že videá patriace do tejto kategórie síce ľudia odoberajú ale až tak nepozerajú.

Obrázok 43: vizualizácia popularity



Pomer videa a odberu	Počet videí kanálu	Počet odberateľov kanálu	Názov kanálu
0,00041	10010	24200000	1theK (????)
0,00299	39701	13300000	ABC News
0,00180	72179	40200000	ABS-CBN Entertainment
0,00015	3597	24000000	Aditya Movies
0,00049	12605	25800000	Aditya Music
0,00014	2492	17800000	Ali-A
0,00007	1581	21400000	America's Got Talent
0,00110	15687	14300000	And TV
0,00002	237	10300000	AsapSCIENCE
0,00006	831	12900000	ATHLEAN-X™
0,00007	981	13300000	Atlantic Records
0,00015	3033	20200000	AuthenticGames
0,00004	466	12500000	Awakening Music
0,00172	21540	12500000	BBC
0,00011	1230	11400000	BBC Earth
0,00055	6978	12700000	BBC News
0,00012	1737	14400000	BBC News Hindi
0,00018	1772	9700000	BIBI
0,00002	310	20200000	Brave Wilderness
0,00010	1794	18600000	Britain's Got Talent
0,00022	4467	20400000	BuzzFeedVideo
0,00036	4035	11300000	CaptainSparklez
0,00001	53	10300000	Clean Bandit
0,00981	138304	14100000	CNN
0,00002	263	12500000	Coke Studio
0,00001	218	23400000	Coldplay
0,00021	3044	14500000	CollegeHumor
0,00034	3772	11100000	Comedy Central
0,00044			

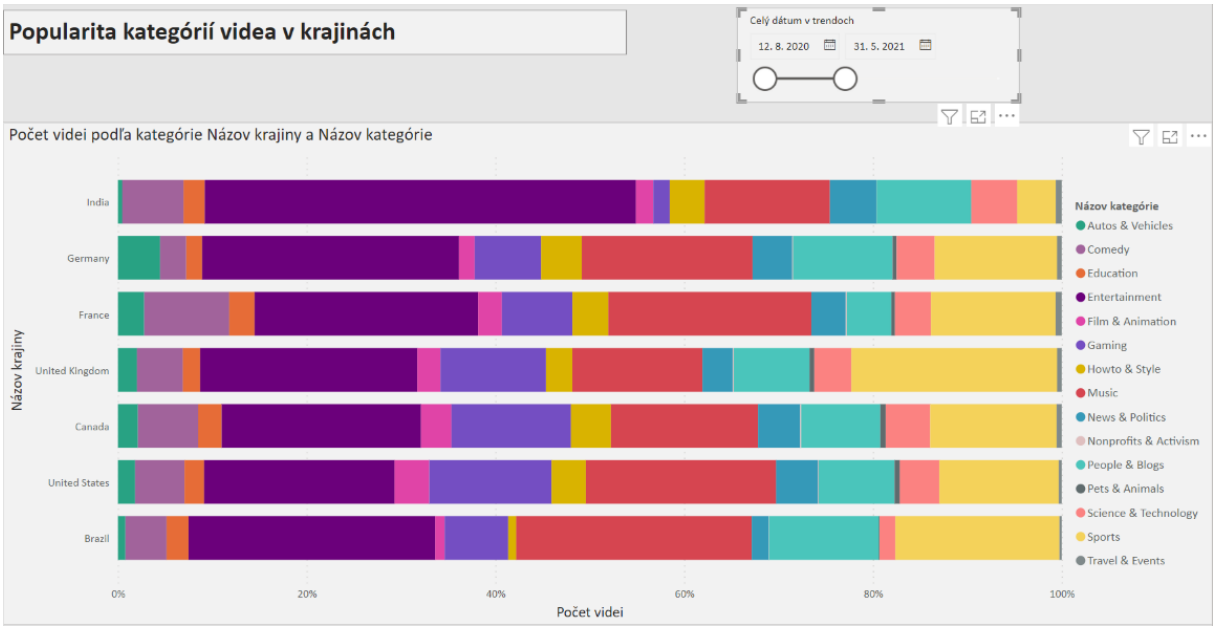
Podľa nášho názoru dôvodom takého veľkého odchýlenia kategórie vzdelávania od ostatných je hlavne fenomén, ktorý sa v posledných rokoch nazýva „prokrastinácia“. Ide o neustále odkladanie povinností a úloh – v tomto prípade môže ísť aj o odkladanie niečoho, k čomu som sa zaviazal (odber kanálu) ale nedokážem sa prinútiť k splneniu záväzku (pozerať sa na vzdelávacie videá). Túto skutočnosť môžeme prirovnať aj k Novoročnému predsavzatiu, kedy si dotyčný človek kúpi permanentku do posilňovne za účelom zníženia váhy, aj tak však nakoniec necvičí.

Zaujímavým výsledkom nám prišlo stále pomerne vysoké umiestnenie kategórie správ a noviniek. Podľa výsledkov z minulej kapitoly, kedy sa CNN objavila na spodných priečkach, sme pomer odberov a sledovania pre túto kategóriu očakávali nižší.

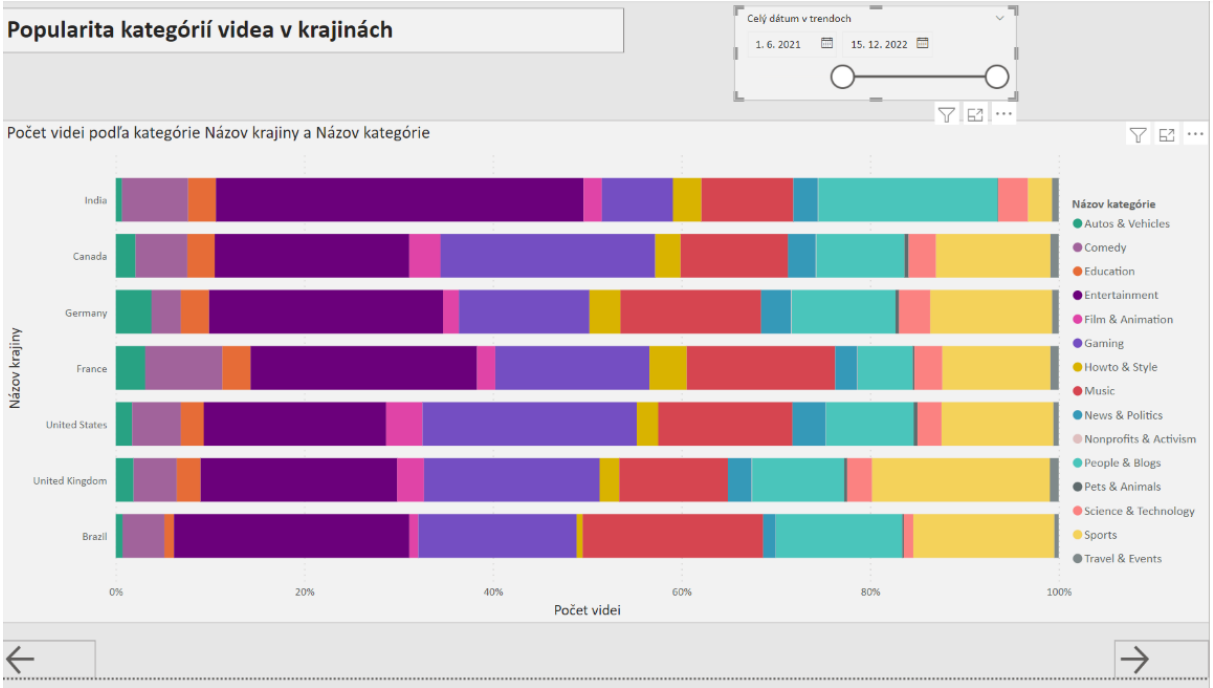
5.7. Zmeny v kategóriách trending videí v čase

Zaujímavé bolo tiež sledovať aj zmeny v kategóriách trending videí v čase. Obmedzený sme však boli rozsahom dát. Ale bolo možné identifikovať tri obdobia, ktoré priniesli veľké zmeny. Prioritou bolo sledovanie obdobie koronavírusu, oproti bežnému obdobiu. Rozsah tohto obdobia bol v našom prípade počiatok dát august 2020, kde už plynulo obdobie koronavírusu a ako koniec sme podľa dát zmien vyhodnotili 31.5.2021. Bolo to obdobie, kedy už poľavilo väčšie ohrozenie a ľudia sa touto problematikou prestali vo veľkej miere zaoberať. Bohužiaľ bežné obdobie nám narušila iná, nie menej nepríjemná udalosť a to začiatok vojny na Ukrajine. Ide teda o tretie sledované obdobie.

Obrázok 44: vizualizácia zastúpenia kategórií v krajinách počas koronavírusových opatrení

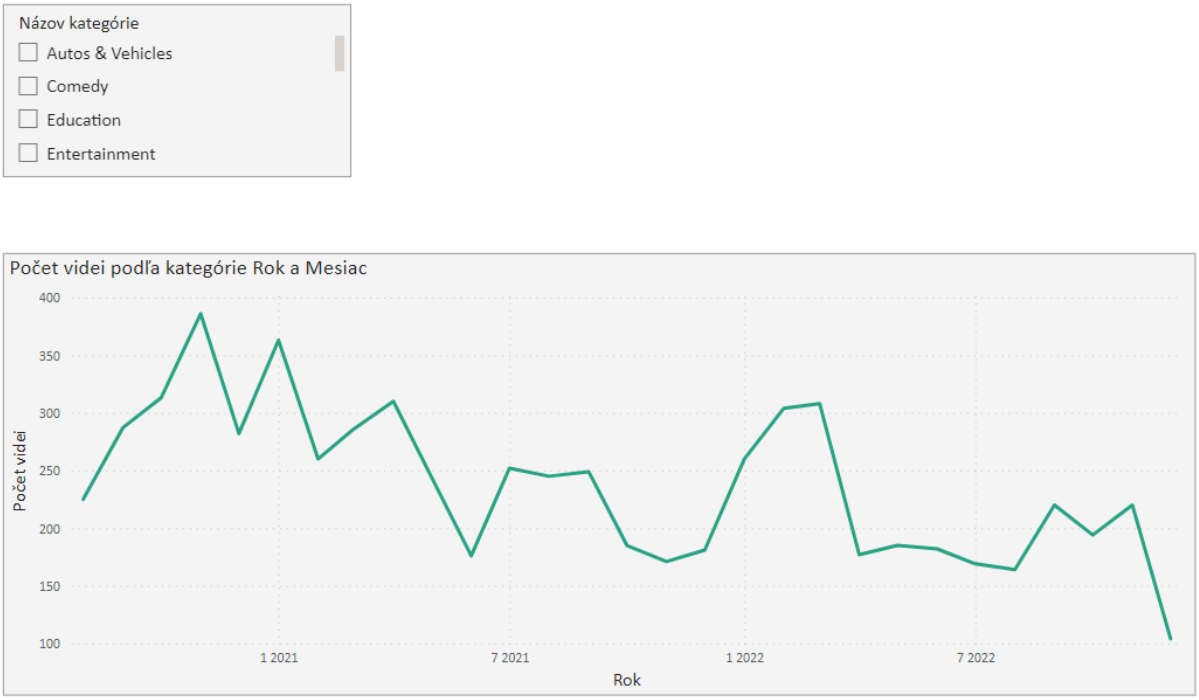


Obrázok 45: vizualizácia zastúpenia kategórií v krajinách po koronavírusových opatreniach



Prekvapivý rozdiel nastal väčšine kategórií, ktoré by sa dali považovať za druhy so zábavným potenciálom. Ide jednoznačne o kategóriu zábavné, ale aj hry. Obe tieto kategórie v období koronavírusu výrazne klesli v trending videách, ale oproti tomu, športové videá naopak vzrástli. Pre vysvetlenie poklesu zábavných videí nebolo identifikované odôvodnenie, keďže sme očakávali spomínaný pojem prokrastinácie v tomto období. V trendoch však pravdepodobne prebila potreba ľudí po informáciách o stave situácie, potrebe samo vzdelávania a športovaniu v domácnosti. Prípadne mierne zvýšené sledovanie kategórie „Howto“, kvôli nutnosti viac sa pohybovať v domácnosti.

Obrázok 46: zmena počtu trenodvých videí označených kategóriou „správy“ v čase

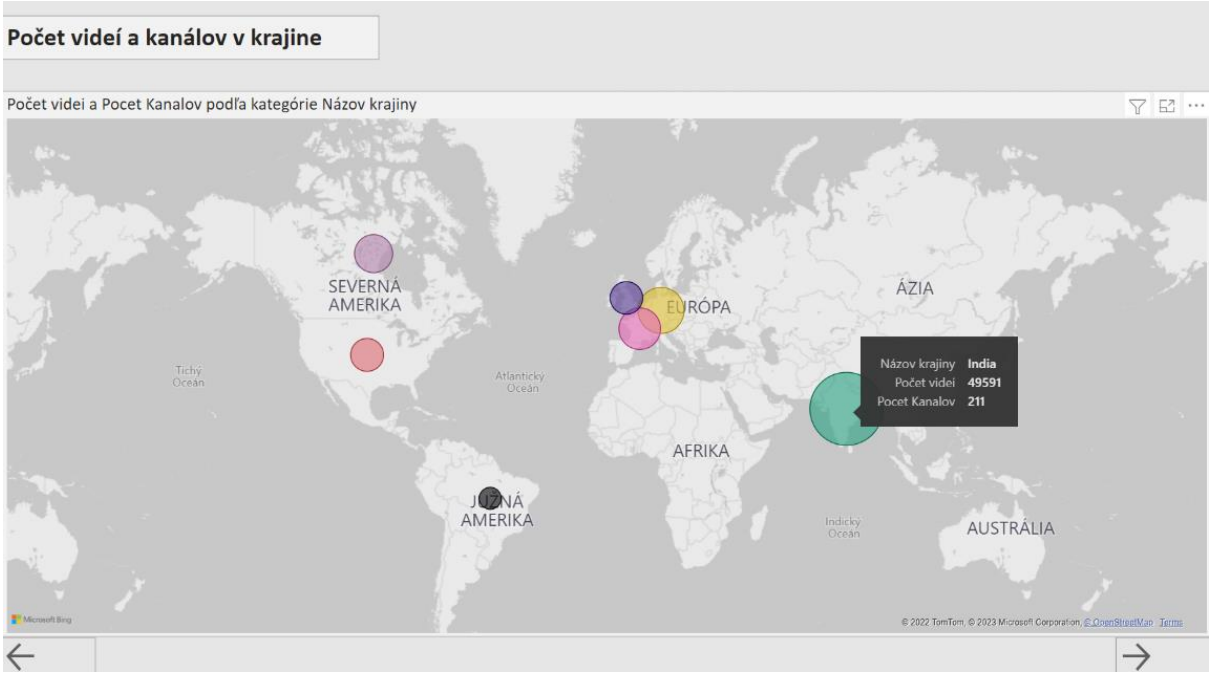


Pri definícií KPI bola stanovená **hypotéza: „Zvýšenie počtu trending videí kategórie správ v období koronavírusu.“**, ktorá už z doterajšieho hodnotenia kategórií bola jednoznačná. Na vyššie zobrazenom grafe, ktorý sleduje počet videí kategórie správ v čase, je pravdepodobne možné dokonca sledovať aj vlny koronavírusu, kde stále vzrastal a klesal záujem sledovať správy. Po postupnom utíchnutí koronavírusu nastal ďalší veľký skok záujmu, ktorý presne zodpovedá udalostiam súvisiacim so začiatkom vojny na Ukrajine. Tento nárazový záujem opäť skokovo klesol a dostali sme sa k pomyselnému bežnému obdobiu. Zmienené pohyby trendových videí jednoznačne **potvrdzujú stanovenú hypotézu**.

5.8. Počet videí a kanálov v krajine

Pre zobrazenie zastúpenia dát s ktorými sme pracovali, sme v Power BI využili vizuál Mapu. Ako môžeme vidieť na obrázku nižšie, najväčší počet videí a kanálov pochádza práve z Indie. Preto sú niektoré výsledky touto skutočnosťou dosť ovplyvnené – ako napríklad spomínané výsledky v kapitole Počet trendových videí pre kanál.

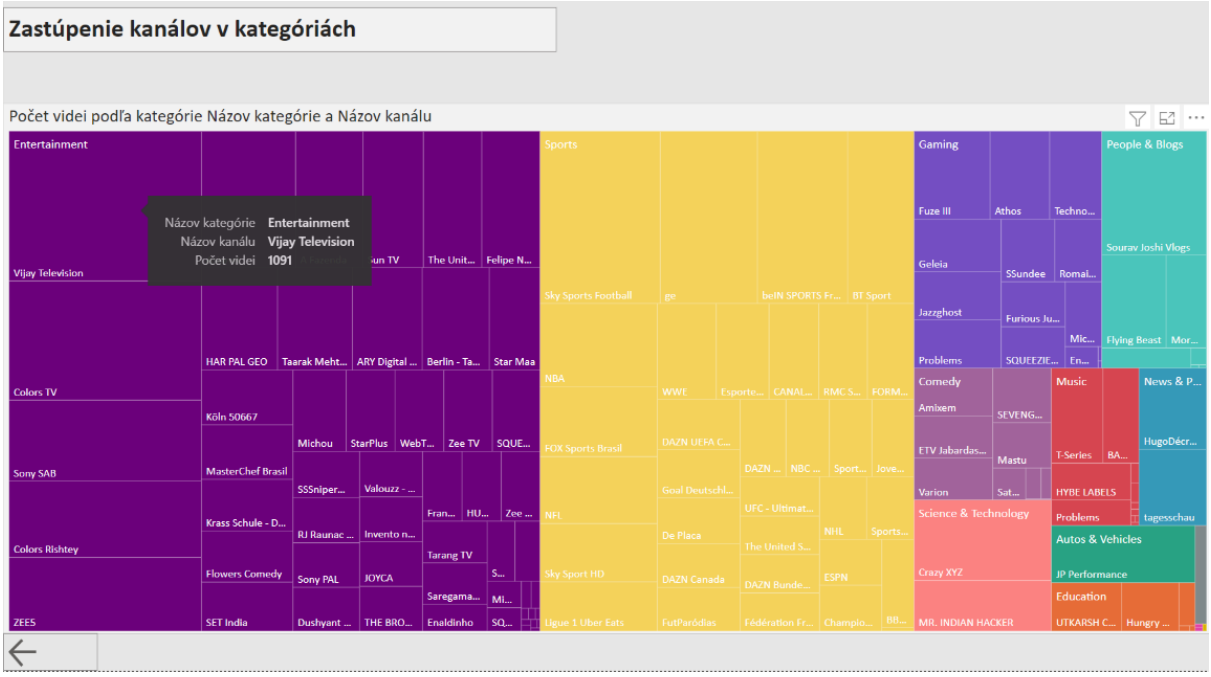
Obrázok 47: vizuál pre zobrazenie zastúpenia počtu trendových vidiei v mape



5.9. Zastúpenie kanálov v kategóriách

Na vizualizáciu zastúpenia kanálov podľa kategórie videa sme využili Treemap. Na obrázku nižšie môžeme vidieť, v ktorej kategórie aký kanál vydáva najviac trendových videí a počet vydaných vide Napríklad Vijay Television v kategórií Zábava.

Obrázok 48: vizuál pre zobrazenie zastúpenia kanálov podľa kategórie videa



6. Diskusia a záver

Samotný výber témy práce pre nás bol ťažkým krokom do neznáma. Voľbu veľmi ovplyvňovala túžba rozumieť aj obyčajným holým dátam a budiť záujem ako v nás tak aj v spolužiakoch.

Téma bola vybraná prevažne pre naše porozumenie a podľa toho, čo ponúkal server Kaggle, ktorý bol naším hlavným zdrojom. Výber však nebol príliš šťastný, pretože sme sa počas vypracovania projektu stretli s mnohými rôznymi problémami. Transformácia dát, automatizácia, úprava datasetov a celkové prehliadanie datasetov, ktoré vo finále ani nemohli byť v práci použité, zabrali viac času ako bolo potrebné. Bohužiaľ táto skutočnosť bola zistená veľmi neskoro.

Aj napriek komplikáciám spojenými so spracovaním dát zo serveru Kaggle však môžeme konštatovať, že sme ich ako tím zvládli prekonať. Každý člen tímu riešil problém svojim spôsobom a dal tak finálnemu projektu široký rozsah možností, ako sa v podobnej situácii zachovať.

Ďalšou veľkou komplikáciou, s ktorou sme sa počas riešenia stretávali, boli problémy s pripojením do školskej siete, prístup na server Treeman.Mendelu.cz, ktorý hosťoval našu databázu. Kvôli tejto nepríjemnosti bol náš tím časovo obmedzený, ale napriek tomu sme sa snažili pracovať s tým, ako to len išlo.

Cieľom našej práce bolo zobrazenie zaujímavostí spojených s dátami o trendových videách a obľúbených kanáloch z dvoch pohľadov.

V prvom sme sa sústredili na získanie výsledkov, ktoré môžu slúžiť záujemcom o kariéru Youtube influencera – čiže Youtubera. Môžu ponúknuť náhľad do „mysle“ jednej z najvýznamnejších webových stránok sveta a môcť tak naštartovať úspešný kanál. Budúci či existujúci tvorca obsahu bude môcť lepšie porozumieť o aký typ videí je asi záujem, na čo by sa mal zamerať a pochopiť, ako približne funguje Youtube algoritmus na určovanie trendy videí a odporúčanie videí ostatným užívateľom.

V druhom pohľade na dáta, sme sa sústredili na zobrazenie zaujímavostí týkajúce sa týchto videí a kanálov v období počas a po koronavírusových opatrení. Odpovedali sme na otázky ako napríklad: „aké ľudia pozerali videá, a aké pozerajú teraz?“, „koľko videí priemerne vydávali kanále a koľko teraz?“ a podobne.

Taktiež sme si v práci vymedzili niekoľko hypotéz, ktoré sa nám podarilo potvrdiť: „počet videí na kanál sa počas koronavírusu zvýšil o viac ako 10 %“ a „zvýšenie počtu trending videí kategórie správ v období koronavírusu.“ a tých ktoré nie: „kategória zábavných videí má viac ako 30% zastúpenie v každej krajine“.

Na záver môžeme povedať, že náročnosť projektu (ako tá kvalitatívna, tak aj časová) nás veľmi prekvapila. Zaiste má celý tím nové skúsenosti a vedomosti, ako nabudúce k takémuto typu projektu pristupovať a čo by sa dalo urobiť oveľa lepšie. Už vieme, čo by sme mali nabudúce očakávať.

7. Literatúra

YOUTUBE OFFICIÁLNÍ BLOG (GOODROW, Cristos), 2017.You know what's cool? A billion hours [online]. [cit. 2022-01-10]. Dostupné z: https://blog.youtube/news-and-events/you-know-whats-cool-billion-hours/

YOUTUBE HELP, 2023. YouTube Partner Program overview & eligibility [online]. [cit. 2022-01-10]. Dostupné z: https://support.google.com/youtube/answer/72851?hl=en

KAGGLE YOUTUBE KANÁL, 2019. What's Kaggle? [online]. [cit. 2022-01-10]. Dostupné z: https://www.youtube.com/watch?v=NzDM Og zsw

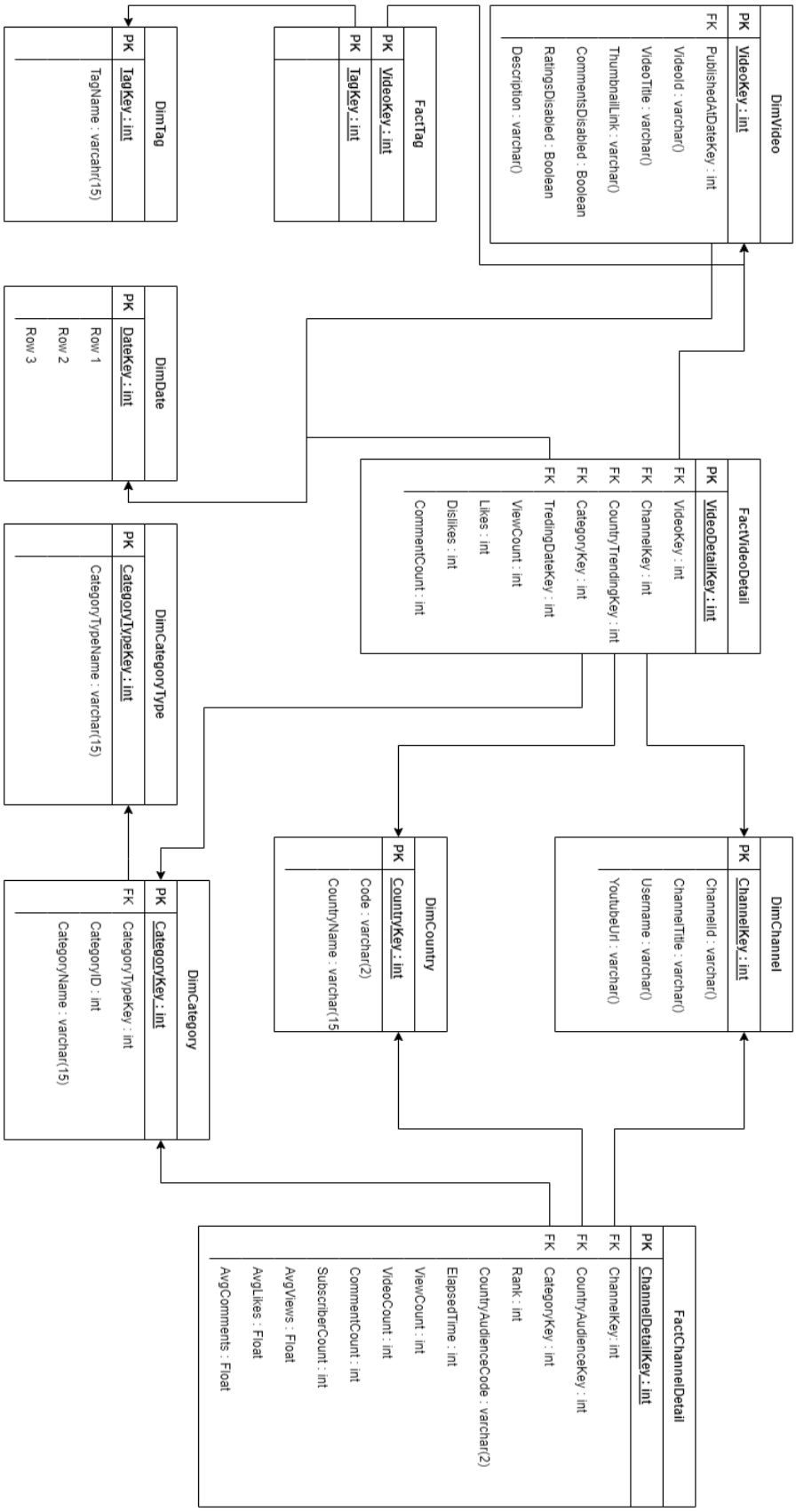
HANKUSOVÁ, Eva, 2020. Klíčové ukazatele výkonnosti (KPI): Co jsou, jak na ně a pár příkladů k tomu [online]. [cit. 2022-01-11]. Dostupné z: https://www.bizztreat.com/blog/klicove-ukazatele-vykonnosti-kpi-co-jsou-jak-na-ne-a-par-prikladu-k-tomu-mnamka

MICROSOFT 365 TÝM, 2019. Klíčové ukazatele výkonnosti (KPI): Co jsou a jak se používají? [online]. [cit. 2022-01-11]. Dostupné z: https://www.microsoft.com/cs-cz/microsoft-365/business-insights-ideas/resources/what-are-kpis-and-how-to-use-them

HOW TO GEEK (MAKVANA, Mahesh), 2022.How to Convert a JSON File to Microsoft Excel [online]. [cit. 2022-01-13]. Dostupné z: https://www.howtogeek.com/775651/how-to-convert-a-json-file-to-microsoft-excel/

ASPSNIPPETS, 2017. Insert all dates between start and end date into table in SQL Server [online]. [cit. 2022-01-12]. Dostupné z: https://www.aspsnippets.com/questions/214738/Insert-all-dates-between-start-and-end-date-into-table-in-SQL-Server/

8. Prílohy


DimVideo
PK VideoKey : int
FK PublishedAtDateKey : int
VideoId : varchar()
VideoTitle : varchar()
ThumbnailLink : varchar()
CommentsDisabled : Boolean
RatingsDisabled : Boolean
Description : varchar()
FactTag
PK VideoKey : int
PK TagKey : int
DimTag
PK TagKey : int
TagName : varcahr(15)
DimDate
PK DateKey : int
Row 1
Row 2
Row 3
FactVideoDetail
PK VideoDetailKey : int
FK VideoKey : int
FK ChannelKey : int
FK CountryTrendingKey : int
FK CategoryKey : int
FK TredingDateKey : int
ViewCount : int
Likes : int
Dislikes : int
CommentCount : int
DimCategoryType
PK CategoryTypeKey : int
CategoryTypeName : varchar(15)
DimCategory
PK CategoryKey : int
FK CategoryTypeKey : int
CategoryID : int
CategoryName : varchar(15)
DimCountry
PK CountryKey : int
Code : varchar(2)
CountryName : varchar(15
DimChannel
PK ChannelKey : int
ChannelId : varchar()
ChannelTitle : varchar()
Username : varchar()
YoutubeUrl : varchar()
FactChannelDetail
PK ChannelDetailKey : int
FK ChannelKey : int
FK CountryAudienceKey : int
FK CategoryKey : int
Rank : int
CountryAudienceCode : varchar(2)
ElapsedTime : int
ViewCount : int
VideoCount : int
CommentCount : int
SubscriberCount : int
AvgViews : Float
AvgLikes : Float
AvgComments : Float